AN ANALYSIS OF EMERGING ETHICAL AND HUMAN RIGHTS ISSUES IN THE HARVESTING OF DATA FROM SOCIAL MEDIA DURING EMERGENCY RESPONSE TO NATURAL HAZARDS

Paul Damien Hayes

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

Trinity College Dublin

2018

DECLARATION

I declare that this thesis has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work.

I agree to deposit this thesis in the University's open access institutional repository or allow the library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

Signed Paul Hayes **Date** 7/5/18

SUMMARY

Digital ICTs can help inform decision-making from a strategic policy-level to tactical response in natural disaster management. This research is primarily concerned with the response phase of natural disaster management. It is interested in discerning how emergency management information systems enhanced by processed social media data can be designed and deployed ethically. To that end, the research is at its base a case study of the output of EU FP7 project Slándáil, a social media powered EMIS developed by international partners across Europe and lead by Trinity College Dublin.

As new technologies can have negative impacts on our moral values and our human rights, this research explores the implications of the Slándáil system (as a gateway into possible uses of the generic technology) under the themes of life, privacy, justice, trust, and responsibility and accountability. This approach is a disclosive analysis, intended to make transparent opaque systems and their implications for moral values. By understanding how the particularities of technology design and use impact moral values, possible adverse impacts can potentially be mitigated.

Data was acquired through semi-structured interviews with key personnel in the Slándáil project, including technologists and emergency managers. Data acquired related to the functionality of the components of the system under development, potential but unimplemented functionality, and on its general potential use and impact on emergency management.

This research utilises a dual theoretical framework of ethical and legal interest. Information Ethics—which was conceived to better address the contemporary problems posed by ICTs—is the first part of this framework. The second is Fiduciary Theory, a constitutional theory asserting that human rights form the blueprints of the state's duties towards its subjects. The use of the dual framework facilitates a comprehensive analysis of the issues at stake, which concern the design of ethical IT systems and the limits of state authority in their deployment.

It is broadly argued that Slándáil-type systems have great potential for supporting moral action and the protection of human rights, but nonetheless risks remain that must be mitigated.

From the perspective of life the studied technology has the raw potential to help avert the tragedy of the Good Will, a situation whereby beneficent agents are unable to avert evil through lack of knowledge or power. The system can help bridge the knowledge and power deficits of emergency mangers and social media users.

The system is argued to have significant impacts on privacy, potentially being implicated in misuse of personal data ranging from its transfer to inappropriate contexts, to indefinite retention. By design the system collects and processes messages and meta-data rich in personal information. It is argued that resort to such systems such be driven by necessity, and data retained only for as long as necessary. Additional technical solutions can be implemented to help preserve privacy.

On the topic of justice; systems that produce information from social media data have the capacity to bias emergency response in favour of internet and social media users, who are likely to be more privileged and resilient than those who are not, and whom (the disadvantaged) it is argued emergency managers should prioritise in natural disaster response. It is argued that emergency managers should use a plurality of information, and that emergency management information systems can be designed to accommodate this.

Trust qualified relations between all agents can be adversely impacted by incorrect information being processed by the system, where the credibility of emergency managers responding to incorrect information, the system which processes it, and social media users that generate it, comes into question. It is argued that automated credibility assessment should be integrated to mitigate this. Function creep is an additional danger, as such systems can be deployed outside of the context of natural disaster management and can be used for broader surveillance and towards oppressive ends. It is argued that such systems should be used exclusively in emergencies, and any expanded functionality should be subject of additional ethical analysis.

Finally, it is argued that the system poses challenges for responsibility and accountability assignment broadly due to the involvement of many agents (both human and artificial) with complex interrelations, varying knowledge, and different interests. Accountability and responsibility can however be supported where roles are clearly defined and known, and where digital record keeping archives information useful in locating faults along the network of agents (for example, user access logs or logs of the system's internal operations).

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

With thanks to my supervisor Dr. Carlo Aldrovandi for giving me the opportunity to embark on this journey. I hope you have found your faith well placed.

Special thanks are due to Dr. Damian Jackson for steadfast mentorship, critical feedback, and patience, which were vital in my efforts.

Thanks are due to Professor Khurshid Ahmad and all participants in the Slándáil project, at the management/administration, end-user, and technical level, for accommodating my research. They are too many to name in full, but each has my respect and gratitude.

With thanks to Aideen Woods for helping with any administrative issues along the way. Also due thanks is Dr. Gillian Wylie for being helpful and supportive during her time as head of school, and beyond.

With thanks to Tomás Kelly, whose mentorship prior to this research prepared me for the demands of research at PhD level.

With thanks to the remarkable John Doyle, my Grandfather, for providing a poor student a roof over his head for his first year of research.

Finally, thanks to my partner Anastasia Rosa Papathanaki for her unwavering patience and support, without which I may well not have completed this research. Grazie. Ευχαριστώ. Σε αγαπώ.

The research leading to these results has received funding from the European community's Seventh Framework Programme under grant agreement No. 607691 (SLANDAIL). The materials presented and views expressed here are the responsibility of the author only. The EU Commission takes no responsibility for any use made of the information set out.

ABBREVIATIONS

AA—Artificial Agent

ANT — Actor-Network Theory

API — Application Programming Interface

CCTV — Closed-circuit television

CI — Contextual Integrity

CID — Criminal Investigation Department

DD — Digital Divide

DHS—Department of Homeland Security

DM — Distributed Morality

DoW — Description of Work

ECHR — European Convention on Human Rights

ECtHR — European Court of Human Rights

EMIS — Emergency Management Information system or EIS

FP7 — 7th Framework Programme

EULA — End-user license agreement

FEMA — Federal Emergency Management Agency

GIS — Geographic Information Systems

GoA — Gradient of Abstractions

GPS — Global Positioning System

HA — Human Agent

HDR — Human Development Report

ICCPR — International Covenant on Civil and Political Rights

ICESCR — International Covenant on Economic, Social and Cultural Rights

ICT — Information and Communication Technologies

IE — Information Ethics

IFRC — International Federation of Red Cross and Red Crescent

IHRL — International Human Rights Law

INFAI— Institut für Angewandte Informatik

IMASH — Information Management System for Hurricane Disasters

IT — Information Technology

LoA — Level of Abstraction

MAS — Multi-agent System

NER — Named Entity Recognition

NIMS — National Incident Management Systems

NLP — Natural Language processing

NSA — National Security Agency

NYCFD — New York City Fire Department

OECD — Organisation for Economic Co-operation and Development

PI — Philosophy of Information

PPI — Public Personal Information

PSNI — Police Service Northern Ireland

RALC — Restricted Access/Limited Control

RPT — Resource, Product and Target

SARS — Severe Acute Respiratory Syndrome

SMS — Short Message Service

SSMM — Slándáil Social Media Monitor

TCD — Trinity College Dublin

UDHR — Universal Declaration of Human Rights

UHP — Ushahidi Haiti Project

UNDP — United Nations Development Programme

UNISDR — United Nations Office of Disaster Risk Reduction

VPN — Virtual Private Network

1 INTRODUCTION

1.1 Introduction

A primary motivation for the present research is the convergence of two distinct categories of challenges to our human dignity and welfare. One such category is that of natural disasters, which threaten our safety and livelihoods and can profoundly disrupt the ordinary day-to-day functioning of communities. The other category of challenges is those posed by the evolution of information and communication technologies (ICTs), which have complex implications for our moral and societal values. The convergence of these categories is represented by the ongoing research and development of technologies that can harvest information from social media sources during times of emergency and process it into actionable intelligence to support the efforts of emergency managers. The field of potential challenges to human dignity posed by ICTs could be expanded to the disaster affected, whose dignity is already challenged.

Challenges however, are merely problems (or puzzles) that can often be solved, and the application of the aforementioned technology to the disaster response context also presents opportunities—primarily information that can assist in saving life and property. The fundamental challenges in the deployment of such technology are to mitigate the potential harms it can do to the very people it is being designed to protect, whilst respecting the authority of emergency managers, that is, not placing them under unreasonable restrictions that would impede their response efforts. What this research will seek to do is engage with the challenges and suggest possible solutions through inquiry into ethical theory and the theory and practice of human rights. Upon achieving this it will be possible to suggest how such technology can be designed and used whilst respecting the dignity of the public and the agency of emergency management actors.

The present researcher was involved in the European Union Seventh Framework Programme (EU FP7) funded project spearheaded by Trinity College Dublin, Slándáil. This project sought to create a system that can ethically harvest, process and analyse data from social media sources during natural disasters. The Slándáil project will serve as a case study in this research. Access to the technologists and emergency management actors enabled the researcher to explore the capabilities and potentials of the relevant technology to interfere with human dignity, as well as protect it.

In his role within the Slándáil project, the researcher had the opportunity to participate in the process of the development of an ethical framework for harvesting social media data for emergency response, a document which served as a project deliverable. During this process, the researcher was exposed to issues of ethical and human rights concerns, which the ethical framework addresses in some depth. This research will broadly revisit these issues, with the benefit of a longer time horizon for more rigorous exploration of the issues and through the lens of different theoretical frameworks, allowing analysis of the issues from new perspectives. The initial ethical framework relied upon the theoretical frameworks of Value Pluralism and States of Exception, whilst this research will rely upon Information Ethics and the Fiduciary Theory of Human Rights. These choices will be explained and defended in Chapter 2.

It is envisioned that this research, exploring and applying contemporary ethical and human rights theories—ripe for further exploration and development—to a new problem will provide an original and novel contribution to academic knowledge. This research is also envisioned to be of practical utility for software developers and emergency managers. Guidelines for the design and use of social media powered emergency management information systems will be extrapolated from the research and will be provided in Chapter 9.

This chapter serves two main purposes. The first of which is to familiarise the reader with concepts and issues relevant to this research—that is, to provide sufficient background context so that they may appreciate and understand the analysis going forward. It will demonstrate the growing threat of natural hazards, as exacerbated by human influenced climate change, and will demonstrate how modern ICTs and information intensive services such as social media can be drawn upon to aid in natural disaster response. This chapter will also take the time to clarify a definition of natural disasters that can be used throughout, and argue that the process of natural disaster management should be framed as one of Human Security (therefore the dignity of the human beings it seeks to protect should be a paramount objective).

Acknowledging that utilising technologies such as Slándáil in natural disaster management has the potential to have adverse implications for the rights and dignity of human beings, the second purpose of this chapter will be to briefly introduce the objectives of this research, and its structure and methodology.

1.2 Emergency Management, Climate Change and Natural Disasters

Risks posed by environmental hazards necessitate emergency management in order to secure the safety and wellbeing of the public. It seeks to reduce the vulnerability of communities to hazards as well as improve coping capacity (Federal Emergency Management Agency, no date). Emergency management consists of several categories of activities including mitigation, preparedness, response and recovery (Phillips, Neal and Webb, 2011, pp. 37–38). This research will be concerned primarily with activities that occur within emergency response to natural disaster.

This section will familiarise the reader with the concept of the emergency-disaster-catastrophe continuum, climate change, and its relationship with natural disasters, with the goal of providing the context of the challenges faced by modern emergency managers. This section will conclude by exploring whether emergency management should be framed as security policy, and what the implications are of this for those affected by disaster.[1]

1.2.1 Disasters and the Emergency-Disaster-Catastrophe Continuum

Perspectives on what precisely constitutes a disaster can vary. Before proceeding it is instructive to consider these different perspectives and unpack an accurate and appropriate definition of disaster that can be used throughout the following.

Two major international organisations, the United Nations Office of Disaster Risk Reduction (UNISDR) and International Federation of Red Cross and Red Crescent Societies (IFRC) define disaster with near symmetry, but with some notable differences. The following is the definition of disaster used by UNISDR (2017):

> A serious disruption of the functioning of a community or a society involving widespread human, material, economic or environmental losses and impacts, which exceeds the ability of the affected community or society to cope using its own resources.

In slight contrast, the IFRC (no date) defines disaster as follows:

[1] It should be noted that while emergency management can be initiated at almost any level of society, by both public and private actors, when this research refers to emergency management and response it is in reference to activities carried out by agents that are instruments of the state.

> A disaster is a sudden, calamitous event that seriously disrupts the functioning of a community or society and causes human, material, and economic or environmental losses that exceed the community's or society's ability to cope using its own resources. Though often caused by nature, disasters can have human origins.

Additionally, the IFRC (no date) notes that disaster is a function of vulnerability, "[t]he combination of hazards, vulnerability and inability to reduce the potential negative consequences of risk results in disaster."

The common thread in both definitions is the emphasis on disruption and the requirement of outside assistance in response. The latter definition emphasises the destructive capacity of disasters, invoking more severe language such as "calamitous."

In the UN Development Programme's (1994, p.29) 1994 Human Development Report, adopts a quantitative perspective of disaster, it is "... an event that has killed at least ten people, or affected at least 100." This definition differs radically from those previously offered, simplistically eschewing concepts of vulnerability and coping capacity of affected communities. An advantage of this quantitative approach is it removes ambiguity by offering two alternative or complimentary criteria—10 deaths or 100 affected—but it can be considered reductive in ignoring the wider dynamics and set of interactions that lead to and result in disaster.

Turning to a scholarly source for the definition of disaster, Charles Fritz (1961, p. 655), as quoted by Philips *et al.* (2011, p.32) defined disaster as:

> ...actual or threatened accidental or uncontrollable events that are concentrated in time and space, in which a society, or a relatively self-sufficient subdivision of society undergoes severe danger, and incurs such losses to its members and physical appurtenances that the social structure is disrupted and the fulfilment of all or some of the essential functions of the society, or its subdivision, is prevented.

This definition differs in its stipulation that a mere threat of disaster can in itself constitute a disaster, on the provision that it causes disruption. The disaster may not be an event, but the perception that an event will occur. The common feature uniting all definitions here is the prerequisite of social disruption. As noted by Philips *et al.* (2011), this definition has three core components: disasters are social events that must impact people, they must cause social disruption for a specific group of people, and outside help for the affected area must be required.

By reviewing these definitions, a succinct and useful definition can be synthesised. For the purposes of this research, with respect to a variety of interpretations from legitimate sources, a disaster will be defined as an event, actual or threatened, that significantly impacts a society or subdivision of that society, causes significant social disruption, and necessitates intervention from outside of that society or its subdivision. A quantitative element to a definition of disaster may be desirable, but is ultimately unnecessary and risks being arbitrary. This definition serves a useful analytical purpose, providing important context to what follows when disaster is discussed, and emphasises that disaster is largely a function of vulnerability, which will be a key point addressed later.

A *natural* disaster arises when a natural agent or hazard of meteorological, climatological, hydrological or geophysical (or even extraterrestrial, consider asteroids etc.) origin is the agent that instigates the disaster. Natural disasters may also trigger technological disasters: consider the Japanese earthquake and tsunami in 2011 that resulted in a nuclear crisis in Fukushima (Phillips, Neal and Webb, 2011). Additionally, natural disasters may be "compounding", where one disaster facilitates another, for example an earthquake causing landslides (Phillips, Neal and Webb, 2011, pp. 115–116).

Disasters exist along a continuum that consists of emergency, disaster and catastrophe (Phillips, Neal and Webb, 2011).[2] Emergencies are more mundane events that can be anticipated and responded to locally (Phillips, Neal and Webb, 2011, p. 34).[3] In the case of a catastrophe, the consequences of the event are particularly severe—most of an area's buildings and infrastructure are impacted or destroyed, outside help is impeded by the situation and a large-scale response is necessitated (Phillips, Neal and Webb, 2011, pp. 35–36).

1.2.2 Climate Change and the Global Impacts of Natural Disaster

Natural disasters are phenomena that are far reaching and have left nary a nation untouched by their sometimes cataclysmic impacts.[4] They have caused immense financial, material and human loss: between 2000 and 2012 natural disasters have

[2] An important note: this terminology use is rooted in the discipline of emergency management and this discussion serves to convey the different levels of severity of disaster situations. In what follows, emergency and disaster will be used synonymously in reference to serious events.
[3] Though they may still occasionally require outside assistance.
[4] See (Munich Re, 2015, p. 2) for a stark infographic illustrating the distribution of geophysical, meteorological, hydrological and climatological events across the world in 2014.

caused an immense 1.2 trillion USD in damage, affected 2.9 million people and resulted in the deaths of 1.2 million (United Nations Office of Disaster Risk Reduction, 2013).

Natural hazards and disasters have had serious deleterious effects on human welfare and have caused incredible material and financial loss. They present a serious threat to the continuity of regular life in many regions. The impacts of some natural weather events are likely to increase going forward, a result of which may well be increased death tolls, more affected persons and greater material costs.[5]

Between 1980 and 2011 climatological, meteorological and hydrological events have been numerous; in total during this time frame there were 3,455 floods, 2,689 storms, 470 droughts and 395 extreme temperatures (United Nations Office of Disaster Risk Reduction, 2012). Recorded Floods and extreme temperatures in particular have seen an exponential increase in frequency since 1980 and 1999 respectively (United Nations Office of Disaster Risk Reduction, 2012).

There is a large body of evidence that global climate has been influenced by human activities, specifically activities that result in anthropogenic greenhouse gas emissions, including primarily carbon dioxide which is produced from burning fossil fuels; a by-product of many industrial processes and combustion based transport (van Aalst, 2006, pp. 5–6). The result of these emissions is reportedly higher atmospheric temperature, or global warming, which is projected to continue into the future (van Aalst, 2006, p. 7).

A probable effect of global warming will be a changes in extreme climate phenomena, potentially including higher maximum temperatures, greater frequency and intensity of precipitation events, greater forest fire and drought risks along mid-latitude continental interiors, increased intensity of tropical cyclone peak wind, an increase in intensity of mid-latitude storms and increased risk of flood and drought damage magnitude in "...temperate and tropical Asia" (van Aalst, 2006, p. 9). The overall consequences of these changes, their likelihoods varying, may be increased risk to human life, as well as the built and natural environments (van Aalst, 2006, p. 9).

In addition to the risks posed by climate change, systems theory in emergency management examines how disasters can arise as a production of interactions of the built (infrastructure and buildings), physical (natural environment) and human components of environment—where one environmental component does not work well

[5] Both due to climate change and increasing populations in urban centres.

disasters may occur (Phillips, Neal and Webb, 2011, p. 45). Expanding industrialisation and population may expose humans to more hazards, both natural and technological (both simultaneously or in succession) (Phillips, Neal and Webb, 2011, p. 67). Evidence has borne that disasters are becoming increasingly costly and affecting greater numbers of people as a function of industrialisation, environmental change to the physical environment and demographic trends (Phillips, Neal and Webb, 2011, p. 98).

The available data offers powerful motivation for actors in emergency management to dedicate adequate resources to all stages of disaster and emergency management. Ironically, the same rapid industrialisation that is conducive to production of greenhouse gases and artificial hazards also facilitates technological advancement, and with that new tools and methods are being created which can be committed to natural disaster management and adaptation to the environmental factors that we as humans are at least partly responsible for making more dangerous. Attention will now be turned to advances in technology that have resulted in an abundance of information, information that can be used to shape disaster management policy, and most pertinently in the context of this research, disaster response.

1.2.3 Framing Emergency Management: What Kind of Security?

Emergency management is concerned with safeguarding security—that is, the security of persons within a state's territory. Security as a term can be amorphous, and the activities that it entails may differ by public policy positions or between different state agents (consider the military versus police activities in pursuing security). Conflations may arise between concepts of general public safety and national security (Zack, 2010, pp. 89–90).[6]

When we think of security measures we might be tempted to think of measures which protect us against external threats, particularly those by human agents (Zack, 2010, p. 91). The debates on security measures become dominated by discussions of liberty versus security, and the appropriate balance that should be achieved in pursuing security. To some extent, at least beyond a certain threshold, the liberty versus security debate may be a false dichotomy. A state that abuses its powers in the implementation of egregious security measures is in itself a threat to the security of its citizens.

[6] Emblematic of this conflation perhaps, for example, was the absorption of the US Federal Emergency Management Agency into the Department of Homeland Security, an organisation founded under the aegis of preventing terrorist attacks (Zack, 2010, pp. 89–90).

Human security as a concept broadly refers to "freedom from fear and want" (King and Murray, 2001, p. 585). It is a concept with broad appeal as it is not rooted in the idea of defence against an enemy, but the protection of human wellbeing—the term incorporates the issues of development and human rights (Futamura, Hobson and Turner, 2011; Hobson, Bacon and Cameron, 2014). In the 1994 United Nations Human Development Report that brought the concept to the fore, it is described as an "integrative" concept, that is centred on people, and which owes much to the notions of human capability and safety (United Nations Development Programme, 1994, p. 24). It shifts discussion from territorial security to comprehensive safety (United Nations Development Programme, 1994, p. 24).[7]

Emergency management should be viewed as one aspect of a process of safeguarding human security, that is, ensuring the safety of persons in the environment in which they are situated. The debate can now be moved from a binary security versus liberty one to one which recognises that the observance of human rights is essential in safeguarding security too—just as natural disasters are a threat to human wellbeing, the state too can be a threat when it enforces policies that have negative implications for personal and political security. Human Security is a holistic concept in which to frame emergency management, it refocuses the discussion on the state's duty to preserve and enhance the well-being of its subjects not only from natural disasters and external threats, but acknowledges that state policy itself can be a threat to Human Security.

1.3 Social Media and the Global Information Society

We are living in an age of information abundance: more data has been accumulated since the commoditisation of computers than in our entire history that preceded the rise of digital technology (Floridi, 2014). Before this age of near ubiquitous computing, humanity produced 12 exabytes of data, a figure which grew to 180 exabytes by 2006 and ballooned to over 1,600 exabtyes between 2006 and 2011 (Floridi, 2014, p. 13). The exponential increases will proceed into the future, as the quoted figure is expected to grow fourfold almost every three years as the digital treasure trove continues to be filled with data (Floridi, 2014, p. 13).

[7] The UNDP (1994) lists the seven major categories of security constituting human security including economic security, food security, health security, environmental security, personal security, community security and political security.

This unprecedented growth of information is of course facilitated by developments in technology that support (if not enhance) the information life cycle.[8]

ICTs are growing increasingly complex and powerful. Floridi (2014, p. 7) explains that Moore's Law proposes that "...over the period of development of digital computers, the number of transistors on integrated circuits doubles approximately every two years". Even as technology grows more powerful and complex, computing devices are becoming more affordable and therefore accessible. Citing research by the Hamilton Project (2011), Floridi (2014, p. 7) indicates that the cost of a device with the iPad2's computing power would have cost over $100 trillion (USD) in the 1950s.[9]

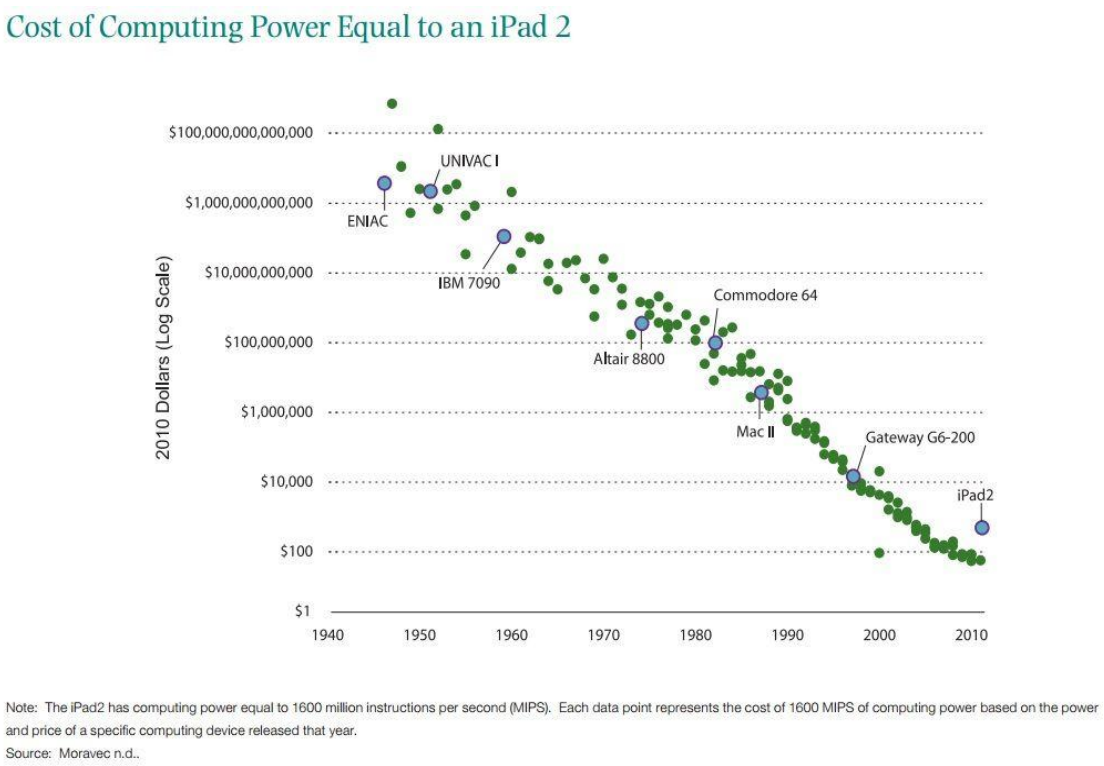

Figure 1: The Cost of Computing Power Equal to an iPad2 (Source: The Hamilton Project at the Brookings Institution, 2011)

[8] Luciano Floridi (2014, p.6) defines the typical informational lifecycle as constituting phases of:

> ...occurrence (discovering, designing, authoring etc.), recording, transmission (networking, distributing, accessing, retrieving, etc.), processing (collecting, validating, merging, modifying, organizing, indexing, classifying, filtering, updating, sorting, storing, etc.), and usage (monitoring, modelling, analysing, explaining, planning, forecasting, decision-making, instructing, educating, learning, playing, etc.).

[9] View Figure 1 to see a graph of The Hamilton Project's research, which deftly illustrates both the evolution of computing technology over time and its increasing accessibility in monetary terms.

Computing technology is powerful and relatively accessible, and using the example of the iPad or smartphones, also highly portable. Computing devices exist as nodes in intangible networks, enmeshed in an interconnected world that has become extremely dependent on them. The number of "connected devices" per person is projected to grow "...from 0.08 in 2003, to 1.84 in 2010, to 3.47 in 2015, to 6.58 in 2020" (Floridi, 2014, p. 12).

The advances in digital technology and resulting information abundance represent an immense opportunity for those involved in natural disaster management—there is now a wealth of information readily available that can potentially be utilised across all phases of disaster management, data that can be accessed in many cases on demand. Census information often exists in online databases that can give important insights into demographic information, and can be especially useful when combined with Geographic Information Systems (GIS). Meteorological, hydrological, climatological (and more) data can be used to build models, projections and simulations. The possibilities stemming from data prevalence are manifold.

The types and sources of information that fill databases and other storage media around the world are diverse. Many of these connected devices are attributable to persons using smartphones and home computers or laptops who are also social media users, and using their devices they generate multimedia content (text, video, images) that is disseminated through social media.

Social media has been defined as the "... tools that enable open online exchange of information through conversation and interaction" (Paquette and Yates, 2011, p. 6). There are numerous categories of social media, which typically provide platforms for the creation, sharing and discussion of video, text or image content including platforms such as video hosting site Youtube, image sharing site Flickr, social networking site Facebook and microblogging services such as Twitter (Starbird *et al.*, 2010; Paquette and Yates, 2011, p. 2).

Social media is becoming a ubiquitous part of life in the 21st century for those with internet access, particularly social networks, and their popularity and reach has been growing. Facebook at time of writing, for instance, has over 1.7 billion active users and membership has been growing exponentially every year since 2008 (Protalinski, 2014;

Statista, 2017b). Also benefitting from a large and mostly growing user-base is Twitter, which as of Q2 2017 had 328 million active users (Statista, 2017c).[10]

In 2017 there were an estimated 2.46 billion social media users, and this is projected to grow to 3.02 billion by 2021, a staggering number that will account for in the region of 40% of the world's projected population by that time, indicating a huge number of persons who can potentially be mobilised or reached out to in the event of natural disasters (Statista, 2017d; worldometers, 2017).

These social media users can generate content at any time so long as they have access to the internet.[11] This means that they can possibly produce information relating to events occurring in their immediate environment expediently, information that can be an excellent asset to emergency managers. In effect, social media users in emergency situations can be relied on as citizen reporters or sources of intelligence, utilising their devices to generate multi-media reports on the situations unfolding around them. The following subsection will examine patterns of social media user-behaviour during natural disasters before providing an overview of current research on how helpful information can be extracted from social media, before then examining the likely future applications of similar research.

1.3.1 Social Media and Natural Disasters

Utilising social media during times of crisis is becoming something of a mainstream response by many actors, both in and out of situations of natural disaster. The Red Cross document *The Case for Integrating Crisis Response with Social Media* outlines numerous examples of persons using social media to signal distress and need—from two young Australian girls trapped in a well who updated their Facebook statuses to cry for help, to an Atlanta city local politician tweeting a call for help upon meeting an unconscious woman on the street (his phone battery was too low to make a 911 call) (American Red Cross, 2010b). These anecdotes do not represent outlying behaviours or attitudes. An

[10] These large numbers of users are not restricted to nations with advanced economies; India boasts the largest Facebook user base (241 million at July 2017), and after the United States (240 million), Brazil (139 million) and Indonesia (126 million) come third and fourth in largest user bases by country (Statista, 2017a). There are also social networks/media that, although they cater for audiences in more specific geographical areas, also have large numbers of users such as China's Sina Weibo (more than 500 million users) and Japan's Mixi (a comparatively smaller but still formidable number of 27.1 million as of March 2012), thus indicating a diverse supply of services beyond the usual social media juggernauts popular across Europe and America (mixi, no date; BBC, 2014).

[11] Or in Twitter's case, any mobile phone coverage is technically sufficient as it offers SMS functionality.

online survey of 1,058 adults conducted by the American Red Cross in 2010 revealed that in the event that 911 was busy in an area-wide emergency, 18% of respondents would try to reach the appropriate emergency service through digital media (American Red Cross, 2010a).[12] This poll also made other significant findings (American Red Cross, 2010a):

- Nearly half of respondents stated that they would use social media to let loved ones know that they are safe
- 69% believed that emergency responders should monitor their websites and social media in order to promptly respond to posted requests
- Three out of four would expect help to arrive within an hour of sending the request over social media

Other behaviours include more general information seeking and sending relating to the disaster. These attitudes and behaviours are understandable and social media would appear to be a very useful communication tool in times of crisis, enabling rapid and multi-directional (one-way, two-way and interactive information exchange etc.) communication—messages can be injected into the social media sphere for consumption and used with great immediacy (Latonero and Shklovski, 2011, p. 6).[13] Social media users benefit from speed and a built in audience, particularly where their communications are publicly viewable.

Legacy media has not been abandoned and still plays an important role in the information life cycle during emergencies, and the information sources or communication tools to which people turn will depend on availability and preference.[14] Telecommunications infrastructure may be critically damaged in an extreme natural event, thereby restricting or disrupting access to social media services (Jennex, 2012). One study for example found that only 19.5% of persons (in their sample population)

[12] Beating walk or drive at 14% and text message at 4%—phone or self-phone dominated at 42% (American Red Cross, 2010a)

[13] Consider that within two minutes of the previously cited Japanese Earthquake of 2011 Twitter users responded (Doan, Vo and Collier, 2011).

[14] In terms of information seeking, research indicates, once again in the case of the Japanese Earthquake, in terms of information seeking behaviour, the old media (TV and radio) were still relied upon sources of information, particularly for non-users of social media (Peary, Shaw and Takeuchi, 2012).

had access to the internet in the aftermath of the Japanese Earthquake (Aizu, 2011, p. 1).

Social media has become an integrated part of the day-to-day lives of many people, and a conventional method of self-expression and communication of information both mundane and important (depending on the context and audience). Seeking and disseminating information during crises is not particularly unconventional for civilians, and indications are that there is a demand for it to be taken seriously by emergency responders. Social media are not the only media type, and legacy media remain valuable sources of information too. One should consider also that old and new media can intermix, whereby information obtained from old media can be reported and disseminated throughout social media (vice versa).

Despite the mainstreaming of social media into society, the type of communication occurring online through social media in crises has been referred to as "backchannel" communications, which are irregular or unofficial communications that are viewed with weariness by public officials due to the potential for misinformation (Sutton, Palen and Shklovsk, 2008, p. 2). However, there is a growing recognition of the benefits of engaging with social media by emergency managers, which will be further addressed later. As social media engagement by emergency management professionals and other statutory agencies increases, and their presence becomes more visible online, the days of this medium being considered "backchannel" may be dwindling.

The information that flows through the sphere of social media during emergencies has the potential to be of great value to emergency managers, and efforts to integrate it into emergency response have the potential to yield great boons. In the following subsection, a closer overview of how people use social media during natural disasters will be offered.

1.3.2 How People Use Social Media During Natural Disasters

Research has identified four broad categories of social media (Twitter communications specifically in the cited article) communications during emergencies. Information can be self-generated by users about a crisis; information can be retweeted (information shared as a carbon copy, essentially); emergency managers send information to affected communities or the general public at official or unofficial capacities; and emergency

managers monitor Twitter feeds to gather information during emergencies (Latonero and Shklovski, 2011, p. 3).

To overview the nature of each broad type of communication, it will be instructive to analyse each of them as separate sub-headings, with slight adjustment.

1.3.2.1 Self-Generated Crisis Messages

In their research on Twitter usage during the 2009 Red River flooding, Kate Starbird *et al*. (2010, p. 6) discuss generative information production, which they state "...is at the core of the information production cycle, providing the raw material that later production behaviour works to shape into a meaningful informational resource." Starbird *et al.* (2010, p.6) code these tweets as original, and note that it occurs in two forms: auto-biographical narrative ("...first-person observations and status updates") and the introduction of common knowledge and adaptation of information from other sources, which they exemplify with the following tweet:

> Thinking that the Red River is not cresting, it's more of a temporary shrinking affect due to the cold weather.

Original tweets accounted for a small proportion of the dataset with which they were working (10%), though locals and peripherals accounted for over 80% of these tweets (Starbird *et al.*, 2010, p. 6).

At critical points during the floods, Starbird *et al*. (2010, p. 7) observe that local Twitter users began to tweet more (for some, nearly exclusively) about flood related issues, resulting in mentions of sand-bagging and importantly, evacuation information. Information that was generated included documenting experiences in sand-bagging in a flood affected city, and dissemination of municipal and flood level updates (Starbird *et al.*, 2010).

In Japan, persons from the most deeply affected areas were most reliant on social media for information, and were also the source of help requests, reports and warnings (Peary, Shaw and Takeuchi, 2012, p. 14; Umihara and Nishikitani, 2013, p. 12).

Another type of information generation is synthetic information production, which is information introduced to the social media space but taken from a plurality of sources—in this case Twitter users essentially extract externally sourced information and compose it in a Twitter friendly format (Starbird *et al.*, 2010). The following quote is an example of synthetic information production (Starbird *et al.*, 2010, p. 7):

> WDAZ says the predicted crest of the Red River is now 52 feet. Follow @egffloodstage to get hourly updates of the river level.

Information generated by individuals may not consist simply of observations and relevant updates. And likewise information is not just offered but is also sought. In the aftermath of the Japanese Earthquake, the types of information offered and sought generally included disaster information, safety confirmation, fundraising, infrastructure status, housing provision, goods provision, moral support, resource saving, volunteer recruitment and special needs support—information behaviours which are evidently typical in disaster/emergency situations and examples of some or all of these can be seen in the aftermath of the Haiti Earthquake and Superstorm Sandy (Ranghieri and Ishiwatari, 2014, p. 138).

It should be clear that a great variety of different types of information arise during crises, and not all of it will be useable for the purposes of emergency response. Messages can be characterised as *signals*—pertinent information—and *noise*—information that cannot be used. Intuitively, the most useful information (signals) is likely to come from local social media users at the scene of an event, who can offer firsthand accounts of the situation.

1.3.2.2 Retweets

Derivative information is also produced on Twitter, and in Starbird *et al.'s* (2010, p.7) research accounted for over 75% of their total sample. Retweets are a primary example of derivative information, it is a convention by which Twitter users share another user's Tweet on their own stream (Facebook and other social networks have similar message sharing capabilities). This acts as a recommendation system (Starbird *et al.*, 2010, p. 7). Research has found that the most retweeted sources in crises are "...mainstream media (especially local media), service organisations, or accounts whose explicit purpose was to cover the emergency event" (Starbird and Palen, 2010, p. 5).

Two types of information are predominantly retweeted; information of broad appeal and information of local utility (Starbird and Palen, 2010). Information of broad appeal is retweeted by persons not directly affected by a crisis event (Starbird and Palen, 2010). Starbird and Palen (2010) found from their Red River and Oklahoma Fire samples that many retweets were prayer requests and few came from locals. The most popular retweet came from a non-local, and contained a weblink to photographs of the Red River floods (Starbird and Palen, 2010, p. 7). Retweets of local utility have more

relevance for local responders and population, obviously carrying more localised information (Starbird and Palen, 2010). Retweets in the Red River event contained information about "...sandbagging coordination efforts, road closures, and river levels" (Starbird and Palen, 2010, p. 7). Retweets pertaining to the Oklahoma Fires contained "[s]helter information (human and pet), fire lines, and first person observations of the emergency..." (Starbird and Palen, 2010, p. 7).

Starbird and Palen (2010) argue that focusing on retweets reduces noise during the collection and analysis of data in real time, and that locally sourced retweets of information are likely to contain the most relevant information. Separate research also notes that informative tweets are more likely to be retweeted than uninformative tweets (Parilla-Ferrer, Fernandez Jr. and Balena, 2014, p. 66).

1.3.2.3 Communications from Emergency Managers/Public Officials

To outline the use of social media by public officials and emergency managers during a natural disaster it is instructive to consider social media communications during Hurricane Sandy, on which a wealth of research exists.

During Hurricane Sandy, government representatives took an active role in disseminating information and engaging with the public throughout social media including the Massachusetts Emergency Management Agency; New Jersey Mayor, Cory Booker; Governors Andrew Como and Chris Christy; New York City Fire Department and many more (Virtual Social Media Working Group and DHS First Responders Group, 2013, p. 17). Such accounts relayed evacuation orders, updates and other information as well as confirming information provided by the public (Virtual Social Media Working Group and DHS First Responders Group, 2013).[15] The City of New York sent over 2,000 tweets, had a Facebook page reach of 322,338 people, gained 176,010 new followers across all social media and notably, the Governor responded to 275 questions and 311 requests, donation and volunteer opportunities on Twitter (Virtual Social Media Working Group and DHS First Responders Group, 2013, pp. 34–35).

The paper, *Online Public Communications by Police & Fire Services during the 2012 Hurricane Sandy* by Amanda Hughes *et al.* (2014) offers significant insight into how public agencies used social media during the course of the disaster response. Most

[15] In New York City, since Hurricane Irene, a Social Media Emergency Protocol has been in place in an effort to ensure a harmonised and managed response to social media communication during disasters (Virtual Social Media Working Group and DHS First Responders Group, 2013, p. 32).

departments in the study area held a Facebook account or website (81%) though much fewer had a Twitter account (13%) (Hughes *et al.*, 2014, p. 3). 70% of fire departments and 60% of police departments had Facebook accounts (Hughes *et al.*, 2014). 25% of departments used Facebook as a communication medium for storm-specific communications with the public, and 7% Twitter (Hughes *et al.*, 2014). The content of communications included information on closures, references to other official sources of information, weather updates and safety instructions (Hughes *et al.*, 2014). 39% of departments that used Facebook to communicate Sandy-related information replied directly to the public, and only 10% of departments did so over Twitter (and the researchers note, sparingly at that) (Hughes *et al.*, 2014).

1.3.2.4 Information Gathering by Emergency Managers

In the article, *Emergency Management, Twitter, and Social Media Evangelism*, the researchers use a case study approach to investigate engagement with social media by emergency management professionals, namely in the Los Angeles Fire Department (Latonero and Shklovski, 2011). They theorise that technological innovation and adoption of social media by fire departments is a result of "evangelists" within the ranks who make concentrated pushes for such adoption (Latonero and Shklovski, 2011). In their study, they found that innovative—if rudimentary—methods were adopted to monitor social media content for emergency response (Latonero and Shklovski, 2011). One of their interview subjects, in describing the monitoring of social media content said (Latonero and Shklovski, 2011, p. 10):

> We're using the new media to monitor, not just send our stuff out via Twitter, but monitor what other people are sending via micro-messaging services, what other people are sending pictures of, what their queries, what their questions are in real time.

In describing how information is validated, a subject states (Latonero and Shklovski, 2011, p. 11):

> I don't have any training, but I use Yahoo Pipes … I dump all my stuff in there, Feed Rinse, all those tools, grind them up and spit them out, and if enough people inside a 20 kilometer area are saying, OMG, or OMFG, that draws my attention. If then I have a traditional media RSS source that says the word, death, explosion, I have a whole algorithm. And then, if it gets good enough, it will make my phone beep. It has to be really--I had a lot of false alarms. My wife wasn't too happy … the phone would buzz all night long, because somebody said something. But people will do certain things, and it lends some degree of credence as to where you want to look closer.

The quoted interviewee also confirmed that multiple sources of information were consulted to verify veracity of reports arising from social media spaces (Latonero and Shklovski, 2011).

Although not spearheaded by state agencies, one of the more notable uses of social media was in the Ushahidi Haiti Project (UHP) in the aftermath of the Haiti Earthquake.[16] This was a volunteer lead crisis mapping effort whereupon volunteers in Boston drew from a plurality of sources such as "SMS, Web, Email, Radio, Phone, Twitter, Facebook, Television, List-serves, Live streams, and Situation Reports" (Morrow *et al.*, 2011, p. 8; International Federation of Red Cross and Red Crescent Societies, 2013, p. 54). It should be noted that importance of SMS[17] as a source was paramount given the low level of internet penetration in Haiti, therefore the contribution of social media to Usahidi's efforts should not be overstated (Morrow *et al.*, 2011; International Federation of Red Cross and Red Crescent Societies, 2013). The work of UHP supported "...situational awareness for strategic, operational and tactical organizations", it was used in conjunction with other sources by the relevant actors (Morrow *et al.*, 2011; International Federation of Red Cross and Red Crescent Societies, 2013, p. 164).

In the aftermath of the Japanese earthquake and tsunami in 2011, a variation of Usahidi, Sinsai.info, also utilised social media content in crisis mapping (International Federation of Red Cross and Red Crescent Societies, 2013, p. 54). In contrast to Haiti, where SMS messages were a much drawn upon source of information (likely as a function of there being few internet users in Haiti), Sinasi volunteers drew primarily on information sourced from Twitter, mapping more than 12,000 reports from the social media site (International Federation of Red Cross and Red Crescent Societies, 2013, p. 54). Sinasi

[16] The utilisation and the mainstreaming of social media into emergency response is evident and not just from conventional state agencies. Communication through and monitoring of social media has been embedded into the American Red Cross' work, for example, through the American Red Cross Digital Operations Center (International Federation of Red Cross and Red Crescent Societies, 2013, p. 60):

> ... DigiDOC synthesizes 'big data' social conversations into situational awareness and, often, anticipatory awareness. It allows Social media posts from the disaster-affected area to be tracked and integrated into response decision-making.... trained digital volunteers work remotely to engage with affected people, providing information, real-time tips, resources, comfort and confidence via social media tools. By routing requests for assistance received through social media to the disaster relief operation on the ground, the centre has opened up an easy-to-use channel for affected populations to communicate directly with the American Red Cross.

[17] Persons could text the short-code 4636 with relevant information that could be processed by UHP (International Federation of Red Cross and Red Crescent Societies, 2013).

mapped requests for assistance from survivors at hospitals and nursing homes (International Federation of Red Cross and Red Crescent Societies, 2013, p. 54).

Services such as these can be of great use to actors in emergency management, but such efforts requiring human volunteers and manual input may soon be rendered obsolete by ongoing advancements in technology, as will be addressed further later.

The current state of monitoring of and data collection from social media by emergency responders would appear to be inconsistent and decentralised, not being embedded in official policy in all instances. Challenges arise from shortages of staff, who are also busily engaged in their traditional roles before the added burden of navigating through the torrents of signals and noise flowing through the sphere of social media (Latonero and Shklovski, 2011, p. 12). It is apparent that utilising social media in emergency response would benefit from improved methods of collecting and processing information. The following subsection will examine the current state-of-the-art in research pertaining to information extraction from social media.

1.3.3 Extracting Information from Social Media during Natural Disasters

Research on the automated extraction of actionable information from social media for emergency response is a burgeoning field producing very interesting and useful technologies and methodologies for effective data collection and processing.

One example of such work is Tweedr. In the development of Tweedr, researchers collected a subset of emergency related tweets and had them annotated by humans, that is classified for relevance (actionable data) and extraction, which involved annotation of damage types (for example, flooding) (Ashktorab *et al.*, 2014, pp. 2–3). This manual extraction can later inform machine learning, allowing automated extraction of signals from noise in social media and the creation of actionable data for emergency responders (Ashktorab *et al.*, 2014, p. 4).

Similar research was conducted using manual classification of tweets and machine learning models, motivated by the fact that "[t]he availability and accessibility of disaster-relevant information can contribute to an effective and efficient disaster response mechanism, which eventually can alleviate damages or loss of life and property during a disaster or crisis" (Parilla-Ferrer, Fernandez Jr. and Balena, 2014, p. 62).

A third example of this research is AIDR: Artificial Intelligence for Disaster Response. AIDR involves the extraction of information from Twitter. AIDR is a "...free software platform that can be run as a web application, or downloaded to create your own instance. It consists of three core components; collector, tagger and trainer" (Imran *et al.*, 2014, p. 3).

Using the Chile Earthquake dataset, the researchers involved were able to extract information on the magnitude of the earthquake and information on, in their own words (Imran *et al.*, 2014, p. 5):

> We could also obtain drilled down numbers about people affected: people dead, people evacuated and people missing. We could also obtain severity of the tsunami warning and the impact distances in various directions.

Research into automated credibility ("offering reasonable grounds to be believed") analysis of tweets has also been conducted (Castillo, Mendoza and Poblete, 2011, p. 675). In research by Castillo, Mendoza, and Poblete (2011), the researchers also trained an automatic classifier in order to determine credibility of information disseminated on Twitter. Their automated classifier was successfully tested, and they found that "...credible news are propagated through authors that have previously written a large number of messages, originate at a single or a few users in the network, and have many re-posts" (Castillo, Mendoza and Poblete, 2011, p. 681). The topic of credibility analysis will be returned to in greater depth in Chapter 7.

This is not an exhaustive list of existing research, and there have been lots of interesting efforts in producing actionable, credible data from social media feeds.[18] Such research is promising for emergency managers who need to obtain structured and accurate information during crises, and can greatly assist them in decision making. Disaster responders variously already use specialised ICTs to aid in emergency management, and incorporation of live data from emergency sites sourced from social media would represent a very useful evolution of such technology, which will briefly be explained in the following subsection.

1.3.4 Emergency Management Information Systems

Emergency management information systems[19] (EMIS or EIS) are any systems that assist in responding to emergency situations (Dorasamy, Raman and Kaliannan, 2013, p. 1835).

[18] The majority of such research uses Twitter, likely because of its comparatively open API.
[19] Examples include:

They are systems which "...should be designed to: support communication during crisis response; enable data and gathering analysis; and support decision-making" (Dorasamy, Raman and Kaliannan, 2013, p. 1835). Such technologies may also incorporate resource management and incident documentation (Thompson *et al.*, 2006, p. 252).

For a more comprehensive insight into the capabilities of EMIS used in a humanitarian context, it is helpful to examine the case of Sahana.

Sahana is an EMIS that was developed in the aftermath of the tsunami that affected Sri Lanka (Currion, Silva and Van de Walle, 2007). Its development was spearheaded by the NGO Lanka Software Foundation and lead by volunteers from the country's IT industry (Currion, Silva and Van de Walle, 2007, p. 63). 40 volunteers contributed, and 100 students contributed by deploying the system and collecting population data (data was collected on 26,000 families) (Currion, Silva and Van de Walle, 2007, p. 63).

In the Sri Lankan case, Sahana's interface consisted of components (which interact through shared databases) such as ..."Organization and People's Registries; the Camp Registry and Request Management System. Subsequently, additional components for inventory management, messaging, situation mapping, and synchronisation have been added to the core set of solutions." Sahana has subsequently been deployed in the "...October 2005 Pakistan earthquake, the 2006 Philippines mudslides, and the 2006 Yogyakarta earthquake in Indonesia" (Currion, Silva and Van de Walle, 2007, p. 63).

In Figure 2, see an example of the online mapping module available on the Sahana website with the precipitation forecast overlay active.

• Sahana Disaster Management Systems for Tsunami (2004), by Sarvodaya.org during Tsunami (2004)
• Information Management System - IMASH for Hurricane Disasters
• Digital Typhoon, a KMS to provide information for typhoons
• PeopleFinder and ShelterFinder
• Strong Angel III (2006), United Nations Development Program
• Tsunami Resource and Result Tracking Systems
• Case Management Systems in Singapore used during SARS (Severe Acute Respiratory Syndrome)
• NIMS (National Incident Management Systems) in USA
• DesInventar System, a historical disaster database and postdisaster damage data collection tool, a project by UNDP and countries such as Latin America, Orissa and South Africa are currently using this system
• Google's Person Finder Tool (launched in 2010) that helped in registering and registering and locating earthquake survivors in Japan (2011), Christchurch (2011) and Haiti (2010) (Dorasamy, Raman and Kaliannan, 2013, p. 1385).

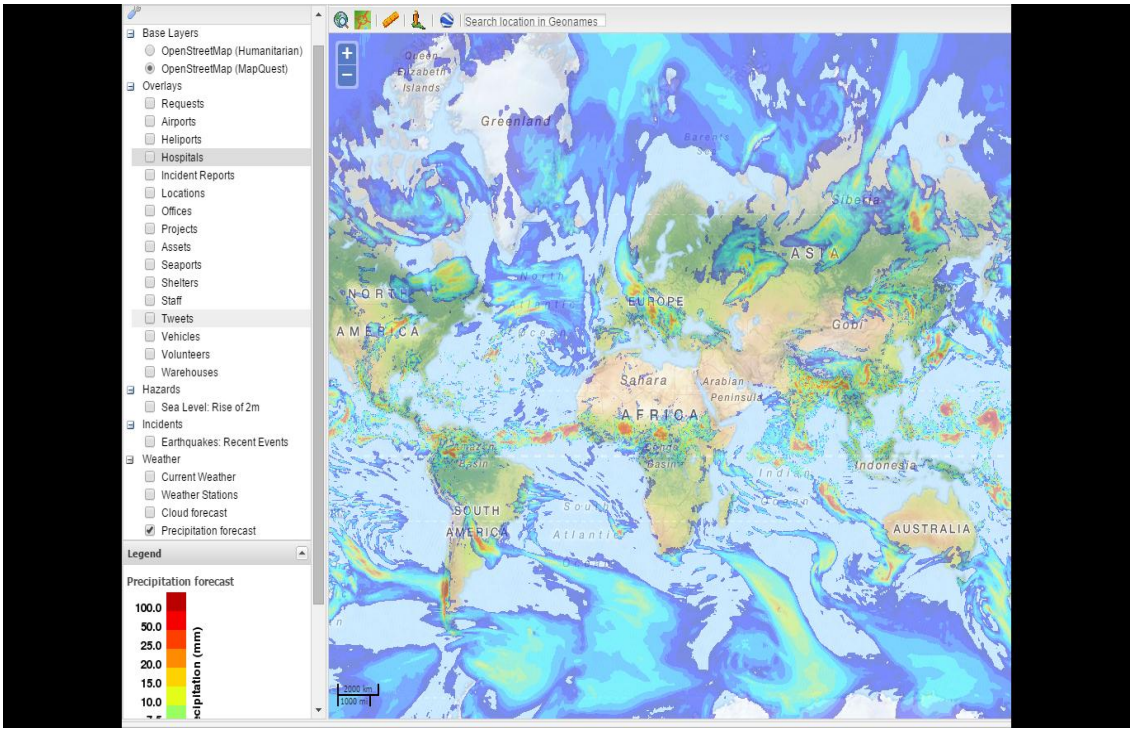

Figure 2: Sahana Online Mapping Module with Precipitation Forecast Overlay (Source: Sahana Foundation, 2015)

Such technologies can be instrumental in aiding coordination and resource allocation, providing instant access to multiple and varying sources of information to aid decision making and help solve difficult problems. Further technological development will see a convergence between EMIS and social media information extraction technologies, it seems inevitable moving forward that future EMIS will include social media modules and mapping overlays illustrating information harvested from social media, which can then be synthesised with other data and ultimately improve the efficacy of emergency response.

1.3.5 Slándáil

Slándáil was an EU FP7 funded project that lead by Trinity College Dublin in collaboration with partners across Europe including academic, business and emergency response actors. Representing the convergence between conventional emergency management information systems and technology that harvests structured information from social media, the project sought to establish a system that ethically harvests relevant information from social media during emergency response to natural hazards (such as floods) that can contribute to situational awareness and provide decision support for emergency responders. The system is a combination of emergency

management software and text/image analytic software.[20] This is a technology under development and will be studied by the researcher to evaluate the particular ethical and human rights implications that may arise from the features and implementation of similar technologies.[21] The research is not intended as an evaluation of the Slándáil project, but will use it as a platform to examine from a close proximity the various technologies under development, their ethical and human rights implications, and the projected use of the final system as well as potential methods to mitigate any adverse ethical and human rights implications for intended beneficiaries or persons whose information is otherwise captured and processed by the technology. It will be used as a model of emerging technology. This will be elaborated upon further in the methodology section.

1.3.6 Data Prevalence: Using Information Responsibly

To return to the concept of Human Security introduced earlier, it would appear that data emerging from social media sources can be applied to guide and improve emergency response and can theoretically be an important feature of protecting Human Security. However, the prevalence of data, its simple availability and ease of access, does not mean that it should be utilised without some consideration for the consequences of its use. Emergency management is an activity protecting Human Security, but information abuse can be viewed as a human right infringement, and would run counter to the goals of Human Security. States and constituting emergency management agencies are faced with the responsibility of using information beneficently—this research will examine how this can be achieved.

1.4 A Value Disclosive Analysis of Social Media Powered EMIS

The utility of harvesting information from social media and effectively isolating signals from noise is intuitively very attractive and poses an excellent opportunity for emergency managers. Nonetheless, it would be foolish to deploy new technologies without attempting to assess their implications for societal values (and human rights). It would simply be irresponsible to deploy technologies intended to benefit humanity without first discerning whether it has adverse implications for societal values, and if so, whether its value threats can be mitigated.

[20] The emergency management software is not dissimilar from the example of Sahana.
[21] At time of writing, the project has concluded and development is currently on pause.

This research then is built primarily on the research questions of "what are the societal value/human rights implications of Slándáil-type systems, and how can value/human rights threats be mitigated?"

Following the work of Philip Brey (2000, 2010), and as will be explored in greater detail in Chapter 3, this research will be styled as a disclosive analysis of the technology developed under the aegis of the Slándáil project. This essentially requires exposing the workings of the studied technology, and examining their impacts on societal values (the manner in which they support them, and the ways in which they may undermine them) (Brey, 2000, 2010). The values selected for analysis were based on perceived relevance based on an early literature review. The value analysis has been restricted to six values. Many more may be impacted, however time and space limitations preclude any additional analysis. The values selected are as follows:

- *Life*—the first value to be explored is that of life itself, which will briefly be analysed in a more optimistic manner in Chapter 4, whereby the potential of Slándáil-type systems to contribute to the protection of life will be outlined.
- *Privacy*—Chapter 5 will explore the implications of Slandáil-type systems for privacy, intuitively a very sensitive value that will be implicated where systems collect and process personal data either incidentally or by design.
- *Justice*—Chapter 6 will explore the implications of Slándáil-type systems for the value of justice with a particular concern for principles of equality and non-discrimination.
- *Trust*—Chapter 7 will explore the implications of such systems for the value of trust, and not only for trust between human beings, but trust between humans and artificial agents.
- *Responsibility* and *Accountability*—The final chapter of the disclosive analysis will deal both with responsibility and accountability, two mutually supportive and often interdependent values of obvious import when concerning the responsible design and deployment of Slándáil-type systems, particularly in view of the difficulties of parsing out responsibility and accountability in complicated webs of agents, both human and artificial.

The disclosive analysis will be supported by a dual theoretical framework of ethical and legal theory; Information Ethics (IE) and Fiduciary Theory, respectively, providing a normative basis for the analysis and conclusions that can be both ethically and legally

persuasive. This selection will be justified and elaborated upon in greater detail in the chapter that follows.

This research is anticipatory (or predictive), endeavouring to detect possible ethical and human rights issues before they manifest with official deployment of such systems as that under study. It will culminate into guidelines for the ethical design and deployment of such systems in Chapter 9, rendering this work timely and potentially very important in providing suggestions for solutions to problems that have yet to manifest and adversely impact relevant stakeholders.

1.5 Conclusion

The purpose of this chapter was to familiarise the reader with the context of this research, including important terminology and concepts. The framing of this research was established from the outset, it is concerned with disaster management as Human Security, that is, the comprehensive protection of the human being from all threats (including internal political threats) and not merely from the threat of violence.

It was demonstrated that natural disasters are a source of great evil in the world, causing untold destruction, and the impacts of natural disaster are likely to grow more severe over time partly as a result of our industrial endeavours. Humankind is partly responsible for the scale of devastation of natural disasters, however with the current developments in ICTs (primarily social media), a new tool has revealed itself that can be used to support natural disaster management and mitigate the evils of natural disaster. The exploitation of technology that can harvest actionable intelligence from the sphere of social media would appear to be something that emergency managers have a duty to capitalise on—responsibility demands that we tap into all available solutions to contemporary problems, perhaps even more urgently where we are indeed responsible for such problems (the impacts of climate change).

This research is motivated by the desire to ensure that technological solutions to locating actionable intelligence on social media can be designed and deployed in a manner that minimally adversely disrupts societal values in order to protect human dignity. As such, in what follows, a disclosive analysis will be conducted that examines the capabilities of a social media powered EMIS (Slándáil), the implications for societal values, and how adverse implications (potentially threats to human dignity) might be mitigated.

2 THEORETICAL FRAMEWORK

2.1 Introduction

The goal of this chapter is to explain and justify the theoretical frameworks chosen for the purposes of analysis in this research. The macroethical theory, Information Ethics (IE), will first be described. This theory is being utilised because traditional ethical theories and approaches face numerous difficulties in view of the complex nature of modern ICTs, whilst Information Ethics offers more intellectual tools to deal with contemporary ethical challenges that they pose. The theory makes controversial propositions, such as granting intrinsic value to, essentially, all instances of reality, and agency to artificial entities. This chapter will take the time to address these important concerns. The pluralistic nature of IE will be addressed, as it is argued that a pluralistic platform is a useful place to begin an analysis of issues of cross-cultural relevance and impact.

The constitutional theory, Fiduciary Theory, will then be described. This is a normative (and descriptive) theory describing and prescribing the nature of the relationship between state and subject, that proposes that human rights are constitutive of a state's obligations towards its subjects. The theory will be outlined, as well as its particular utility within and application to matters of emergency. The particular strengths of the theory will be outlined, such as how it replaces consent with trust, and how it offers a potential bulwark against the normalisation of emergency measures, or the threat of permanent emergency.

Finally, the rationale for the dual framework will be offered with the primary arguments being that two issues are at stake, the design of an ethical ICT system to be used in emergencies and the use of this system by state authorities. IE's ontology is useful for examining the interactions between complex processes and providing the grounds for moral evaluations. Fiduciary Theory is useful for demarcating the limits of state power in emergencies. The ontology of IE can help to evaluate the changing nature of the ethical challenges, and can dialogue with Fiduciary Theory to perhaps make more persuasive arguments about not just what is ethically important, but constitutive of a state's obligations towards its subjects.

2.2 Information Ethics

Information Ethics (IE) is a contemporary macroethical theory that was developed by Oxford scholar Luciano Floridi in order to address the ethical challenges posed by digital ICTs. Floridi (2013) believes that the standard ethical theories are not entirely equipped to deal with these ethical challenges and that viewing information and computer ethics as micro-ethics[22] is also inadequate for addressing the complexities of the relationship between people and digital technologies.

Floridi (2013, p.6) argues that our new technologies are "re-ontologising"; a neologism that refers to the radical transformative effects of technologies that not only re-engineer entities anew, but change their very essence. The example he offers of re-ontologising is one which was covered to some extent in Chapter 1, which is the "...transition from analogue to digital data and then the ever-increasing growth of our informational space" (Floridi, 2013, p. 7). This re-ontologised world, in what is a hyperhistorical[23] age for many, where metaphorical "cyberspace" is becoming synonymous with our "umwelt",[24] calls for an innovative approach to its new challenges (Floridi, 2013, p. 16).

In the informational ontology adopted by IE, we as human beings are, at a high level of abstraction (LoA), information entities ("consistent packets of information") inhabiting an "infosphere" and sharing it with other information entities (Floridi, 2013, p. 65).[25] The infosphere, Floridi argues, can be used synonymously with reality (Floridi, 2014, p. 41). Being, and all that exists as expressions of Being, are of fundamental value—even more basic than life and pain or suffering (Floridi, 2013, p. 16). IE adopts a principle of ontological equality, that is, all entities within the infosphere have a basic intrinsic moral value—everything that exist demands respect and holds "...an initial, overridable, minimal right to exist and develop in a way appropriate to its nature" (Floridi, 2013, pp. 68–69). This may appear counter-intuitive, and it will be addressed in greater detail further in this chapter, however it bears noting that the extension of intrinsic value to

[22] Using resource, product and target—RPT—informational approaches (Floridi, 2013).

[23] As "history" is synonymous with historical record of information, "pre-history" refers to a situation of no recorded information, "history" to a situation of historical record, and "hyperhistorical", another neologism coined by Floridi, refers to our current situation of massive amounts of efficiently managed information (Floridi, 2013, p. 6, 2014a, pp. 1–24). Floridi (2014, 1-24) uses the term as an adverb, and not necessarily in reference to any particular time period, as indeed there are still those, such as uncontacted Amazonian tribes, who live prehistorically.

[24] That is, the distinction between on-line and off is fading (Floridi, 2013).

[25] The infosphere is: "...the totality of Being, hence the environment constituted by the totality of informational entities, including all agents, along with their processes, properties and mutual relations" (Floridi, 2013, p. 16).

instances of reality[26] does not detract from the particular importance and value of human beings. IE adopts the view that while all things, at a basic level of abstraction, have a minimal intrinsic value, they are not all alike in dignity; those agents (humans) which have the capacity to contribute to the wellbeing of the infosphere are unique in their dignity (Floridi, 2013, p. 76).

In IE, the field of agency is extended beyond intentional, conscious and self-reflective entities to other entities,[27] contingent on the LoA at which that entity is observed. At its most basic an agent is "...a system, situated within part of an environment, which initiates a transformation, produces an effect, or exerts power over time", however for the purposes of analysis this LoA is too high, and Floridi (2013, p. 140) provides the following qualities which characterise an agent, which can be artificial or natural:

- Interactivity
- autonomy
- adaptability

Should an action produced by an agent have moral effects, the agent qualifies as a Moral Agent (Floridi, 2013, p. 147).

Note again that the status of agency is contingent on LoA. A level of abstraction is a collection of observables analysed with a particular goal, where everything but the observables relevant to the analytical goal are abstracted (Floridi, 2013, pp. 29-51). A car mechanic repairing a faulty engine for instance, may in the LoA he employs abstract all elements of the vehicle except the engine. At one LoA an entity may qualify as an agent and yet at another it may not. Floridi (2013, p.140) offers the example of a webbot that filters spam email into an appropriate email folder. At one LoA, the basic user's, the webbot will appear as an agent as it demonstrates the requisite characteristics of interactivity, autonomy and adaptability—the bot can learn user preference, autonomously performs its task, and interacts with other objects within its particular environment (Floridi, 2013). When the algorithm by which the webbot "learns" is revealed, perhaps at the programmer's LoA, it is no longer an agent, it loses the adaptability characteristic (Floridi, 2013, p. 145).

[26] In whatever form they take, from a rock to the Mona Lisa.
[27] Ordinarily these are intuitively exclusively human.

The temporospatiality of the method of abstraction is also flexible—Floridi (2013) notes for instance that a corporation can qualify as an agent, however when speaking of complex networks of agents it may be more appropriate to use the term multi-agent system (MAS).

The converse of an agent is a patient, a recipient of an action. A patient can be an entity within the infosphere, or the whole infosphere itself (Floridi, 2013). An agent can be a patient too, or a patient an agent, depending on the LoA or specific circumstances of a case as it may be (Floridi, 2013).

Good and evil in IE are governed by the concepts of entropy[28] and flourishing, derived from which are IE's core normative principles (Floridi, 2013).

Entropy (metaphysical) is defined as any kind of destruction or corruption[29] of entities understood as informational objects, that is, any form of "impoverishment of Being" (Floridi, 2013, p. 67). The concept of flourishing is referential to general welfare, preservation and improvement, to entities and the world as a whole.

Following from this, the core normative principles of IE are (Floridi, 2013, p. 71):

> 0 - entropy ought not to be caused in the infosphere
> 1 - entropy ought to be prevented in the infosphere
> 2 - entropy ought to be removed from the infosphere
> 3 - the flourishing of informational entities as well as of the whole infosphere ought to be promoted by preserving, cultivating, and enriching their well-being

The action of an agent is more praiseworthy the higher on the scale the effects of its action(s) are (Floridi, 2013, p. 71).

Actions may fail to be either good or evil, that is, they may fall below a threshold beyond which they have a meaningful effect, environments may be morally inert or fault tolerant (Floridi, 2013, p. 266). Floridi (2013) argues that many actions are neither good nor evil as their actual effects are insubstantial, however good (or evil) can be derived from the aggregation of many morally negligible actions, uniting separate actions into one larger one that can push beyond an environment's fault tolerance/inertia. Distributed Morality (DM) is the consequence of aggregated actions that produce good

[28] Not to be confused with its application in thermodynamics.
[29] Of corruption, Floridi (2013, p. 67) states: "[c]orruption is to be understood as a form of pollution or depletion of some properties of the entity, which ceases to exist as that entity and begins to exist as a different entity, minus the properties that have been corrupted or eliminated".

(Floridi, 2013, pp. 261–266). DM may require a multi-agent system (MAS), which is "...is a conglomeration of interacting components, known as agents, capable of cooperating to solve problems that typically are beyond the individual capabilities or knowledge of each agent" (Floridi, 2013, p. 104).

Information Ethics is not agent-centred in its analysis of moral action. It is a patient-centred ethics of care concerned with the impact of an action on its recipient (Floridi, 2013). It demands consideration for the effects of a given action on a patient, which can be any informational entity, which is always intrinsically valuable though with varying levels of dignity based, essentially, on its own capacity to contribute to the infosphere's flourishing (Floridi, 2013).

With the central underlying concepts of IE described, it is now instructive to examine the model of moral analysis that IE supports.

2.2.1 Moral Action in the Infosphere

In IE, "...the dynamics of reality are analysed in terms of its information flows" (Floridi, 2013a, p. 68). Floridi (2013, p. 108) provides an informational model of moral action that consists of several different components consisting of:

(1) the agent
(2) the patient
(3) the interactions between the patient and the agent
(4) the agent's general frame of information
(5) the factual information concerning the situation insofar as it is at least partially available to the agent
(6) the general environment in which the agent and the patient are located, and
(7) the specific situation in which the interactions occur

Figure 3 illustrates a model of these components as they exist in a moral situation. The moral situation (or envelope) is the specific region of the infosphere in which a moral action occurs. In the moral situation, A and P are objects representing the agent and patient, the action is a message or process that is directed from agent to patient which can essentially make P better or worse off, have a feedback effect for A, or have ramifications that extend beyond the immediate situation and in turn affect more patients (the "*propagation* of an operation")(Floridi, 2013, p. 109). The action that harms P may be judged as morally bad, it may be an instance of entropy in the infosphere. Referring again to Figure 3, note the shell, which is A's general frame of information or its subjective world, and is defined by Floridi (2013, p. 107) as "...A's

moral values, prejudices, attitudes, likes and dislikes, phobias, emotional indignations, moral beliefs acquired through education, past ethical evaluations, moral experiences..." and "...the shell represents the ethical and epistemic conceptualizing interface between Alice and her environment."

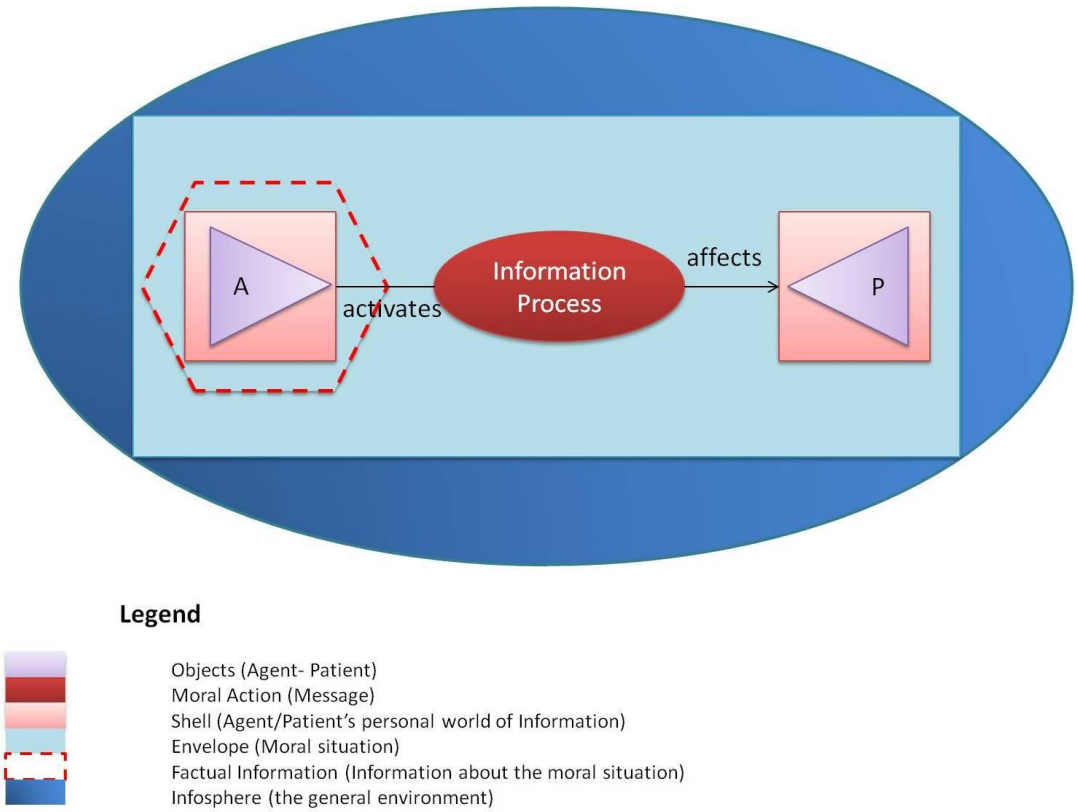

Figure 3: The Informational Model of a Moral Action (Source: adapted from Floridi, 2013, p.108)

As IE extends the field of agency note that A may not be human at all but even an artificial agent. In the case that A is an artificial agent, and not a conscious being, its shell is (intuitively—and not argued by Floridi) more likely to be a reflection of its internal protocols that determine its operations and functions, perhaps a reflection of its human creators' own shell.

Morally responsible agents become more responsible for outcomes in moral situations as their acquisition of factual information about a situation increases (Floridi, 2013). Only humans are morally responsible agents, as they are "...aware of the situation and capable of planning, withholding, and implementing their interactions with the infosphere with some degree of freedom according to their evaluations" (Floridi, 2013, p. 68). Humans can be both responsible and accountable for their actions, however artificial agents cannot be responsible, but can be held accountable in a sense—their

positive or negative contribution to an outcome can be identified, and although they cannot be punished or censured as typically understood, they can be reengineered (Floridi, 2013).

The combination of methods and concepts endorsed by IE allow the morally responsible agent to visualise the network of agents and patients interacting within an ethical scenario at varying levels of abstraction, or at different levels of granularity put more simply. In a world of complex networks, with artificial agents and human agents enmeshed in multi-agent systems, where there is a necessity to examine the ethical impacts of these software agents, and the humans and organisations that design and deploy them, the theory is advantageous in helping to trace and evaluate the impacts of their actions on their recipients, and analysing issues of responsibility and accountability within these networks of complex interactions.

The technology we have at our disposal, that gives us greater power and access to information, that places us in a position to avert evil, makes us more responsible for its prevention and mitigation (consider the combination of technologies that help warn us and respond to the evils of Natural and Heteronymous Agents such as earthquakes)—our modern technological resources are making us responsible for natural and artificial evil (Floridi, 2013). To meet the task and rise to the challenges before us, we are required to maximise the moral impact of our technology—we need to ensure that our software systems do no harm and are consequently designed ethically, and are used ethically in the wider systems in which they operate. This is the thrust of the arguments provided by Floridi— and the basis of Distributed Morality—that to push moral action above a threshold, every component of a system must be ethically designed. To do so when we have the resources is an ethical imperative, to promote flourishing of the infosphere and prevent and remove entropy.

2.2.2 Why Information Ethics?

The reader might ask why Information Ethics and not a more traditional theory such as deontology or consequentialism. The arguments adopted here are simply that traditional theories come under some strain to resolve contemporary ethical challenges involving ICTs. Additionally, in an interconnected world, IE serves as a platform for pluralistic discussion of ethical problems.

2.2.2.1 The Limits of Traditional Theory and Approaches

Firstly, IE is a useful macroethical theory with a unique informational ontology that is more accommodating of the challenges presented by modern ICTs than standard ethical theories.

Floridi (2013, p. 58) argues that the ethical challenges posed by modern ICTs "...strain the conceptual resources of *action-oriented* theories..." in a serious way, resulting in distortions caused by the projection of the standard account of (human) agency onto ICTs or the delegation of responsibility to ICTs as "...increasingly authoritative agents...". Floridi (2013, pp. 58-59) argues that Kantian ethical principles such as the laws of impartiality and universality face difficulty in resolving issues such as ostensibly victimless crimes (for example "...computer crimes against banks"), that is, traditional Kantian theory fails to adequately approach non-anthropocentric contexts. Conceptual distance from the results/consequences of an agent's actions also diminish its sense of moral responsibility and accountability (Floridi, 2013, p. 60). Moral evaluations are challenged by the perceived marginality of an agent's action, the consequences of which may be indirect and unperceived by the agent (Floridi, 2013). Ethical evaluations of actions perceived as marginal or insubstantial, often dealing with intangible assets, are rendered difficult. Consequentalist principles fail as the agent cannot fathom the consequences of their acts (Floridi, 2013). In sum, effective action-oriented analysis of acts in the infosphere (the networked, hybrid environment) are quite severely hampered (Floridi, 2013).

IE is patient-oriented and non-standard, granting all forms of reality (interpreted informationally) a minimal (though not absolute) value. The responsible moral agent assesses their actions not based (necessarily) on compliance to rules or by aggregate happiness produced by their action, but considers the patient in their analysis of a situation—the patient which does not have to be tangible or conscious (Floridi, 2013). IE also emphasises how aggregate courses of action, and not just individual potentially morally negligible actions, impact the infosphere (Floridi, 2013). IE demands that the responsible agent evaluate wider networks of action, and how each individual strand involved contributes to the larger morally charged one. It promotes respect and consideration for the impacts of the action of the individual agent on a patient, and it demands answers to the question of how evil can be mitigated from aggregate action emerging from networks of human and non-human agents, and how good moral actions

can be harnessed from networks of human and non-human agents. It is an ethics that recognises that humans are responsible for what they create, and that they have a responsibility to make them "good" (artificial agents), and demands that humanity contribute to the general wellbeing of their informational environment (Floridi, 2013).

Using IE in the field of information and computer ethics also proves superior to a micro-ethical approach. Floridi (2013) argues that information and computer ethics, as a micro-ethical endeavour that evaluates the ethics of information in three discrete vectors as a resource, a product, and a target (RPT),[30] is also inadequate for properly addressing the problems of our time. The tripartite RPT model is inadequate as it is too simplistic and not inclusive enough—ethical issues can cut across the three information vectors, or arise from interactions between them (Floridi, 2013, pp. 25–26). As a macroethical approach, IE situates the three vectors within the informational environment and instead of viewing information in a semantic sense only, it treats information as objects, entities, or as described, a part of the very fabric of reality (Floridi, 2013, p. 27). Thus, by taking a macroethical approach that adopts an informational ontology, rather than a micro-ethical approach, it is possible to more effectively address the ethical issues that arise from the use of current ICTs. Again, IE is more conducive to the evaluation of complex interactions between entities in the infosphere. IE as a macroethics with its holistic ontology[31] is capable of analysing the full range of issues arising from the flow of information and its impact on the environment and ultimately society.

Figures 4 and 5 illustrate the shift from the external to the internal model of RPT represented by IE.

[30] Information ethics from the resource perspective is essentially "...the study of the moral issues arising from the 'triple A': *availability, accessibility* and *accuracy* of informational resources independently of their format, type and physical support." From the product perspective, "...covers moral issues arising, for example, in the context of *accountability*, *liability*, *libel legislation*, *testimony, plagiarism, advertising, propaganda, misinformation, disinformation, deception*, and more generally of *pragmatic communication*...". As a target, it refers to violations/intrusions within information environment, including issues of privacy, security and hacking (Floridi, 2013, pp. 22-25)

[31] It is worth noting that IE is applicable beyond ICT related scenarios.

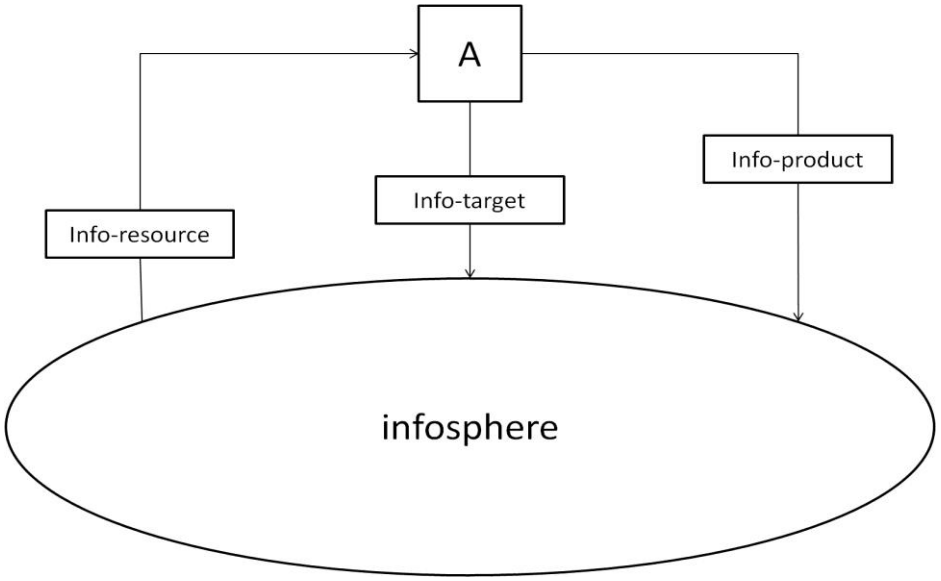

Figure 4: The external Resource, Product, Target Model (Source: adapted from Floridi, 2013, p. 20)

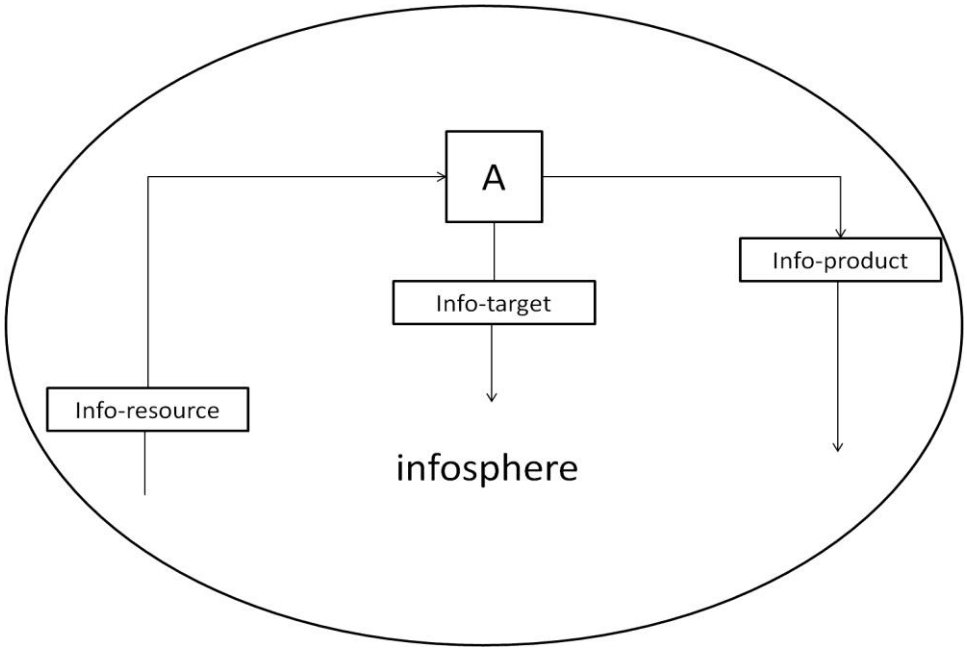

Figure 5: The internal Resource, Product, Target Model (adapted from Floridi, 2013, p.27)

2.2.2.2 Pluralism

In a globalised world of porous borders that supports rapid inter-cultural communication and interactions between connected individuals, an ethical framework that supports a

plurality of views is necessary to overcome what are global challenges that affect us all—transcending cultural and national boundaries. IE was developed as an ethics that supports pluralism without endorsing relativism or any kind of dominant moral ontology (Ess, 2009; Floridi, 2013a). Charles Ess (2009, p. 159) notes that IE and the Philosophy of Information are an important development in providing a common foundation for ethical debate, by marrying elements of western philosophy and eastern traditions—potentially providing a bridge between them both. He argues that it can conjoin norms and values whilst preserving irreducible differences between cultures, that is, it can support common solutions to problems without dismantling differences that exist between cultures (Ess, 2009, pp. 215–226).

Floridi (2013, p. 321) argues that there is no right LoA independent of the goal of analysis, and that IE's "...analysis is based on the reasonable choice of a plausible and fruitful approach to the sorts of new ethical problems emerging in the information society." Floridi (2013) denies that any LoA is a good LoA (it must be appropriate towards an analytical goal); the interpretive, minimally prescriptive nature of IE means that it is not absolutist, that is, there is not only one LoA that should be adopted. IE does ask however that an LoA be deployed towards the overarching goal, in the ethical context, of promoting flourishing in a deeply ecological sense—however this sense of promotion of a common good is a basic value that cuts across culture and positions (Floridi, 2013). IE has the ability, because of its "lite ontology" to tolerate and interface with other "local" ontologies (Floridi, 2013, p. 299).

2.2.3 Popular Objections

The reader familiar with traditional theories might find some of IE's assertions unconvincing. The principle of ontological equality has proven to be one controversial point and the agency of artificial entities another. This chapter will highlight defences to arguments against each in order to assuage the concerns of the reader who might find the claims of IE counterintuitive.

2.2.3.1 The Intrinsic Value of Information Entities

The intrinsic value or moral worth, ascribed to informational entities/objects through its principle of ontological equality is an interesting point of contention and one which, due to its quite radical assertion, warrants a particular defence.

Notable objections to this come from Kenneth Einar Himma (2004) and Philip Brey (2008), who both find the claim of ontological equality untenable for different reasons. Himma (2004, p. 145-159) for his part, finds the concept of abstract informational entities that exist in logical space problematic, even moreso the modelling of human beings as informational entities. He objects to the notion of the human qua information object, something characterised by "...properties, attributes and behaviours" that is ontologically distinct from the physical human being itself (Himma, 2004). He believes that IE cannot offer satisfactory explanations as to why we might respect a stillborn child more than a rock (Himma, 2004, p. 152).

Brey (2008) offers objection to ontological equality based on what he views as a conflation of value and respect, arguing that what Floridi deems worthy of respect might be more deserving of it because of its instrumental or extrinsic value rather than some inherent intrinsic value. IE suffers from an "untenable egalitarianism" (Brey, 2008, p. 112).

The overarching themes of such objections pivot on IE being supererogatory, counterintuitive, untenably egalitarian and suffering from issues of unclear metrics for measuring intrinsic value (Floridi, 2013). Floridi offers arguments against all charges.

IE, as stated earlier, grants entities only "...a *minimal*, *overridable* and *ceteris paribus*," moral value, there is no absolute prohibition against the destruction of anything within the infosphere (Floridi, 2013). Evil is unavoidable, the best we can hope for is to do *more good,* which is our moral imperative (Floridi, 2013). IE invites the responsible moral agent to consider intrinsic value only initially, on the basis that Being is good, and non-Being is evil, that is, there is a basic value to that which constitutes our reality. Sometimes the nature of something that exists, for instance toxic waste, would even necessitate its destruction; for while it constitutes our reality, it also corrupts it. Neither does IE endorse a quantification of intrinsic value, it is based on qualitative assessments and practical wisdom (Floridi, 2013). Floridi (2013, p. 306-326) also argues that assigning intrinsic value to sentient and non-sentient realities is not in itself a novel deviation from a philosophical norm, with parallels to be found in traditions and cultures from Spinoza to Buddhism.

As to the argument of IE failing to explain why we might afford a stillborn child no more or less respect that a rock, the theory does not provide an immediate answer, but

invites the ethicist to do so within its framework. Both the rock and the child have minimal overridable value and warrant respect to begin with, however we must then chose our LoA and the complete set of factors involved. Without resolving the debate, perhaps we respect the child beyond its intrinsic value for what it might have been, or perhaps we respect, and additionally care for it more, because of the grief its death causes its parents, or that the loss of a child (capable of contributing to the infosphere's flourishing in many ways) is especially tragic. Regardless, the child, as a human, has dignity whilst the rock has not, and warrants more respect on closer analysis. Respect is "initial", and "overridable", not absolute. In this manner it might be said that extrinsic or instrumental value is being appealed to. This does not necessarily undermine IE—it might be argued that extrinsic factors can override initial intrinsic value. What is important is that we at least start from a place that acknowledges the inherent value of existence—after this, other factors can be weighed in to an ethical analysis.

All information objects then hold intrinsic value and warrant some respect as instances of reality, on initial analysis, though beyond this initial appeal there is no reason that one cannot necessarily consider extrinsic factors. There is no conflation between value and respect, an information object demands respect initially because of its value as an instance of reality.

2.2.3.2 Agency

The approach taken by IE to agency, as described, reduces an agent to something, which at a given LoA, displays characteristics of autonomy, interactivity and adaptability. Intentionality is not a perquisite. A moral agent is one whose actions are sufficiently morally charged (Floridi, 2013).

This is a non-traditional approach again, and one with which Himma (2009) provides arguments to the contrary. According to Himma (2009), agency is contingent on mental or intentional states.[32] He argues that "...X is an agent if and only if X is capable of performing actions. Action are doings, but not every doing is an action; breathing is something we do, but it does not count as an action" (Himma, 2009, pp. 19–20). Actions require intentional states, or decisions, according to Himma (2009), and are not

[32] He defines moral agency as (Himma, 2009, p. 24):

> ...for all X, X is a moral agent if and only if X is (1) an agent having the capacities for (2) making free choices, (3) deliberating about what one ought to do, and (4) understanding and applying moral rules correctly in paradigm cases.

mindless autonomous processes. Moral agency encompasses accountable beings, beings that are governed by moral standards and are by virtue bestowed with moral obligations—they are evaluated by their moral standards (Himma, 2009, p. 21).

As such, consciousness is a prerequisite for moral agency. Artificial constructions (artefacts) and other entities without the capacity for reasoning and deliberation fail to be moral agents. Entities incapable of mental states, that "do" rather than "act", fail to be agents (Himma, 2009).

Therefore the objections put forward may generally fall into the domains of intentionality, freedom and responsibility/accountability. Floridi (2013) has prepared a defence to each of these objections. Firstly, he argues that intent is not a perquisite for moral agency. Including intent in the analysis presumes privileged access to an agent's intentional/mental states that cannot be guaranteed (Floridi, 2013, p. 149). It represents a problem for analysis and evaluation when it relies upon psychological speculation as opposed to observable information (Floridi, 2013, p. 149). Knowing whether an agent intended its actions is only important when attributing responsibility—the agent played a moral game independently of a conscious state, and should consequently be evaluated as a player in a moral game (Floridi, 2013, p. 149).

Secondly, Floridi (2013) argues that artificial agents are free in the sense that they are non-deterministic systems which are interactive, autonomous, informed and adaptable. They are free to choose their actions to a certain (albeit perhaps limited) extent.

Thirdly, AAs cannot be responsible—this remains the exclusive domain of the human being, however it can be held accountable, just not in a traditional sense of censuring or punishing someone to change their behaviour (Floridi, 2013). AAs can be re-designed, re-engineered or simply destroyed (Floridi, 2013). This is analogous to how accountability works with human beings, and functionally similar. Blame follows responsibility, which rests on the shoulders of humans who may or may not have been in a position to design a system that performed morally, and as such may or may not also be held accountable (Floridi, 2013). The infosphere is a more complicated construct than the old analogue world, and traditional approaches to agency are weaker at identifying and evaluating sources of moral action, especially when one considers the plurality of autonomous and quasi-intelligent entities active in our informational environment—the old approach limits analysis, especially when one considers the

phenomena of Distributed Morality. IE may have changed the goal posts, but justifiably and defensibly so.

2.3 Fiduciary Theory

The crux of Fiduciary Theory is "...that the state and its institutions are fiduciaries of the people subject to state power, and therefore a state's claim to sovereignty, properly understood, relies on its fulfilment of a multifaceted and overarching fiduciary obligation to respect the agency and dignity of the people subject to state power" (Criddle and Fox-Decent, 2009, p. 347). The theory hinges upon the moral concept of dignity, which legal scholars Evan J. Criddle and Evan Fox-Decent (2009, p. 348) argue is not abstract but rooted in the legal relationship between state and subject. Dignity, they argue, "...reflects the intrinsic value of agency..." (Criddle and Fox-Decent, 2009, p. 365).

Examples of fiduciary-type relationships include ..."agent-principle, partner-partnership, joint venturer-joint venture, parent child and guardian-ward" (Criddle and Fox-Decent, 2009, p. 349). The basis of a fiduciary relationship is that one party holds discretionary administrative power over the legal or practical interests of another (the beneficiary), and the beneficiary is vulnerable to the power entrusted to the fiduciary, that is, the beneficiary cannot protect him/herself from an abuse of power from the fiduciary (Criddle and Fox-Decent, 2009). The discretionary administrative power of fiduciaries is defined by three principle criteria (Criddle and Fox-Decent, 2009, p. 349):

- It other-regarding: the power discharged is not self-regarding, it is not explicitly in the interests of the fiduciary but regards those of the beneficiary.
- It is purposive: the power held is limited and discharged towards limited purposes.
- It is institutional: the power held is located within a legally permissible institution.

The fiduciary may hold power to perform an action which the beneficiary can legally do him/herself, or to perform an action that s/he cannot do him/herself, for example, children who are not legal adults (Criddle and Fox-Decent, 2009, p. 350).

The final example is notable, as the Fiduciary Theory draws inspiration from an analogy offered by Kant involving parents and children that lays the conceptual foundations of

the fiduciary relationship (Criddle and Fox-Decent, 2012).[33] In the parent-child relationship; parents bring a child into a condition to which they cannot consent, they unilaterally create a human being that cannot support itself (Criddle and Fox-Decent, 2009, p. 354). The child places its parents under obligation to meet its needs. This right is innate and legal, granted to the child simply for being born (Criddle and Fox-Decent, 2009). The child, as a legal person (and not a thing) cannot be abandoned or destroyed, as it has entitlement to freedom, has intrinsic dignity and must be treated with respect (Criddle and Fox-Decent, 2009). The fiduciary obligation arises as a result of recognising the child's "...moral capacity to put her parents under obligation" (Criddle and Fox-Decent, 2009, p. 354).

The state assumes a fiduciary role over its subjects, its various branches (executive, legislative and judiciary) hold discretionary administrative power over those subject to this power (Criddle and Fox-Decent, 2009, p. 352). The powers wielded by the state are other regarding, purposive and institutional. Those subject to the state's power ("legal subjects") are not, as private parties, entitled to wield the state's powers—they are subject to the state's discretionary administrative powers and therefore vulnerable to it (Criddle and Fox-Decent, 2009, p. 352).

In Fiduciary Theory, state power "...denotes the effective authority of a state to rule and represent a permanent population within a given territory" (Fox-Decent, 2011, p. 90). Sovereignty can be *de facto*, the "...brute ability to govern through effective institutions..." or *de jure*, where state power is legally and politically authorised (Fox-Decent, 2011, p. 90). Whether or not state power is *de jure*, because of the state's

[33] Kant's analogy can help conceptualise the State-subject relationship and the nature of the fiduciary relationship. As argued by Criddle and Fox-Decent (2009, p. 354):

> As persons, children cannot be treated as mere means or objects of their parents' freedom to procreate. Rather, they are beings who by virtue of their moral personhood have dignity, and dignity proscribes regarding them as if they were things. By the same token, legal personality and the idea of dignity intrinsic to it, supplies the moral basis of the beneficiary's right to the fiduciary obligation. A relationship in which the fiduciary has unilateral administrative power over the beneficiary's interests can be understood as a relationship mediated by law only if the fiduciary (like the parent) is precluded from exploiting his position to set unilaterally the terms of the relationship with the beneficiary. The fiduciary principle renders the beneficiary's entrusted interests immune to the fiduciary's appropriation because those interests, in the context of fiduciary relations, are treated as inviolate embodiments of the beneficiary's dignity as a person. In other words, the fiduciary principle authorizes the fiduciary to exercise power on the beneficiary's behalf, but subject to strict limitations arising from the beneficiary's vulnerability to the fiduciary's power and her intrinsic worth as a person.

position of irresistible power over a vulnerable population, it is required to discharge its fiduciary duties. The legitimacy of the former, however, is highly contestable.

The state-subject fiduciary relationship is argued to be a legal and political relationship, and one which has legal consequences (Criddle and Fox-Decent, 2009, pp. 356–357).

Fox-Decent and Criddle (2010, p. 315) frame the overarching fiduciary obligation of states as the establishment of a "...regime of secure and equal freedom under the rule of law," and argue that human rights are "...the blueprints or structure of this regime".

Fox-Decent and Criddle (2010, p. 302) use the Fiduciary Theory to reframe human rights as legal entitlements that are "...grounded in the state subject fiduciary relationship...". They argue that human rights are defined by numerous characteristics under the Fiduciary Theory (Fox-Decent and Criddle, 2010, p. 302):

- They are *relational* and *institutional* in that they are responses to threats that emerge in the relationship between state and subject.
- They are *legal* and *nonpositivist*, constituting the necessary conditions of legal order under Kant's theory of right.
- They are *practical*, seriously regarding rights enshrined in international human rights instruments.
- They are *aspirational* and *universal*, because they are necessary to ensure conditions of secure and equal freedom.
- They are *deliberative*, because they can be refined under democratic deliberation.

The fiduciary interpretation of the state subject relationship, owing to its Kantian roots, is grounded in a principle of non-instrumentalisation, that is, people are to be viewed as and treated as ends, and not means (Fox-Decent and Criddle, 2010, p. 310). Additionally constituting the Fiduciary Theory's normative dimension is the ideal of non-domination, that is, public institutions are duty bound to protect subjects from arbitrary power (Fox-Decent and Criddle, 2010, p. 310).[34] Both non-instrumentalisation and non-domination are tied into the idea of independent agency, and require respect for an individual's capacity for self-determination without undue interference or simply the threat of interference (Fox-Decent and Criddle, 2010, p. 310). Securing conditions of non-

[34] As argued by Fox-Decent and Criddle (2010, p. 310): "...human rights are correlates of the State's duty to secure conditions of noninstrumentalization and nondomination".

instrumentalisation and non-domination, or securing human rights that arise as an obligation of these principles, is a duty impelled by the assumption of sovereign power—when a state fails to discharge its fundamental (fiduciary) duty, it loses its legitimate claim to govern on the behalf of and as representative of its subjects (Fox-Decent and Criddle, 2010). The state is authorised to secure legal order with human rights among its constraints. Not only are human rights constraints to state power, they also constitute its duties (Fox-Decent and Criddle, 2010, p. 315).

There are three desiderata that further build the substance of human rights under Fiduciary Theory emerging from the fiduciary requirement that the state act for the good of its subjects, rather than narrowly the interests of those agents embedded within the state official power structure (Fox-Decent and Criddle, 2010, p. 318):

- *Integrity*—"...human rights must have as their object the good of the legal subject rather than the good of the State's officials."
- *Formal Moral Equality*—"the fiduciary State owes a duty of fairness or evenhandedness to legal subjects because they are separate person's subject to the same fiduciary power. Human Rights therefore, must regard individuals as equal cobeneficiaries of the fiduciary State."
- *Solicitude*—"...human rights must be solicitous of the legal subject's legitimate interests because those interests, like the interests of the child vis-à-vis the parent, are vulnerable to the State's nonconsensual power."

Before proceeding any further, it is important to note the implications the *deliberative* characteristic of human rights. Whilst the Fiduciary Theory impels commitment to non-instrumentalisation and non-domination, in order to secure a regime of secure and equal freedom under the law, emerging threats to dignity and agency may redefine the catalogue of human rights (Fox-Decent and Criddle, 2010, p. 317). That is to say, Fiduciary Theory accommodates the formalisation of human rights that have not yet been recognised as such.

The Fiduciary Theory does not hold that all rights are absolute, and therefore supports the practice of derogations and limitations. States are however bound by *jus cogens* (peremptory) norms, which prohibit instrumentalisation and domination and are consequently non-derogable (Criddle and Fox-Decent, 2009). Recall that rights in

conflict, or rights that come into conflict with legitimate State goals, may be subject to limitation.

Under the Fiduciary Theory, non-absolute rights are considered "presumptively mandatory", however limitations are permissible on the condition of proportionality, and that they are justifiable to the public—the state implementing limitations must accept political and legal responsibility, that is, it can be held accountable to the subjects from which its power flows (Criddle and Fox-Decent, 2009, p. 385).

2.3.1 Fiduciary Theory and Emergency

Fiduciary Theory has the potential to be an exceptional normative framework for state action in the midst of emergency, and can provide clarity on appropriate procedure in an international legal landscape of ambiguity, divergence and contradiction (Criddle and Fox-Decent, 2012, p. 42). Criddle and Fox-Decent (2012, p. 51) refer to the body of international law relating to state power in emergency situations as International Law's Emergency Constitution, and they criticise it on the basis of the contradictory sets of norms and practices that exist between different international instruments and bodies. Criddle and Fox-Decent (2012, p. 51) position Fiduciary Theory as one which can provide a coherent theoretical foundation for international human rights law (IHRL), and one which rebuts the Schmittian argument that "...sovereign discretion displaces legality during national crises."

Criddle and Fox-Decent (2012) examine the implications of emergency through the two-tiered analysis of law regulating entry into states of emergency (*jus ad tumultum*) and law regulating state action within a state of emergency (*jus in tumultu*).

2.3.1.1 Jus ad Tumultum

Under *jus ad tumtultum* a state may declare a state of emergency if the circumstances of a threat impede its ability to provide a regime of secure and equal freedom through ordinary means. The state is duty-bound to prevent the instrumentalisation or domination of its subjects by both its own institutions and private actors, and during times of crisis may need to resort to extraordinary measures that could be contrary to its fiduciary duties in normal circumstances (Criddle and Fox-Decent, 2012).

Criddle and Fox-Decent (2012, p. 48) refer to the European Court of Human Rights (ECtHR) *Lawless v. Ireland* [1961] case and argue that Fiduciary Theory supports three of

the four criteria listed by the ECtHR for the justification of a state of emergency. The three criteria are that circumstances warranting an emergency declaration must be:

- present or imminent
- exceptional
- constitute a threat to the organised life of the community

The rejected criterion is that the circumstances must affect the entire population (Criddle and Fox-Decent, 2012). This criterion is rejected as being incompatible with the state's duty to provide secure equal freedom to all its subjects (Criddle and Fox-Decent, 2012). The state, obligated to provide secure and equal freedom for all its subjects, is required to restore or maintain public order for persons even in limited geographic areas and can use emergency powers strictly necessary towards those ends (Criddle and Fox-Decent, 2012, p. 64).

On the first criterion, states are not authorised to implement emergency measures to combat hypothetical threats—credible evidence should be furnished justifying implemented measures, and proving that such measures are *necessary* to avert a crisis that would disrupt legal order (Criddle and Fox-Decent, 2012). The state has a fiduciary obligation "...to evaluate potential threats cautiously and deliberatively, with appropriate solicitude to those who bear the burden of rights-infringing measures" (Criddle and Fox-Decent, 2012, p. 63).

On the second criterion, emergency powers may be utilised only in the event that normal restrictions are insufficient for the state to discharge its basic overarching fiduciary obligation (Criddle and Fox-Decent, 2012). Exigent circumstances must exist that render traditional laws, procedures and practices inapplicable before that state can use emergency powers (Criddle and Fox-Decent, 2012). The state is authorised to adopt emergency powers for the duration of the emergency, regardless of its duration, but terminate them immediately after the passing of exigent circumstances (Criddle and Fox-Decent, 2012).

Of the third listed criterion, the circumstances that threaten the organised life of a community are so if they "...disrupt the state's ability to guarantee its subjects' secure and equal freedom" (Criddle and Fox-Decent, 2012, p. 61).

A satisfactory framework for the invocation of a state of emergency has been outlined. Next, the extent and limits of state power during emergencies under Fiduciary Theory will be outlined.

2.3.1.2 Jus in Tumultu

In *jus in tumultu* the principle focus of justification for measures taken during emergency is that of necessity (Criddle and Fox-Decent, 2012). Measures taken by states must be such that they are strictly necessary to restore order, and are required to be proportionate to the ends sought, ensuring that measures are no more intrusive than absolutely required by the exigencies of the situation (Criddle and Fox-Decent, 2012). As the Fiduciary Theory supports general limitations outside of emergency, the authors argue that special justifications for derogations should be offered (Criddle and Fox-Decent, 2012). The authors argue that in practice, the Fiduciary Theory would only rarely authorise a state to derogate from human rights treaty provisions that already contain limitation clauses (Criddle and Fox-Decent, 2012, p. 68).

It is essential that states provide details to their subjects of the circumstances of the emergency, the exact rights suspended, the measures taken and the reason for the implementation of those measures (Criddle and Fox-Decent, 2012).

Thus public justification is an important aspect of derogation. The theory additionally requires international notification, however this is done primarily for the benefit of the state's subjects and not the international community as it provides the subjects with the means to contest emergency measures (though it might also be noted that Criddle and Fox-Decent frame international bodies such as the United Nations Security Council as secondary guarantors of human rights) (Criddle and Fox-Decent, 2009, 2012, p. 69).

The combination of public and international notification serves to give individuals subject to the state's power numerous avenues to contest measures taken, including through political and judicial processes and independent human rights commissions (Criddle and Fox-Decent, 2012, p. 70). Contestation is intrinsic to the realisation of non-domination, as subjects require mechanisms to challenge arbitrary uses of power(Criddle and Fox-Decent, 2012).

It bears restating that the Fiduciary Theory proscribes states from restricting or derogating from rights that are recognised as *jus cogens*.

As a final remark, it is also notable that Fiduciary Theory eschews the somewhat consequentialist theoretical model of interest balancing applied to international law that supports rights restrictions upon the weighing of interests (for example, torturing a suspect of kidnapping to obtain a victim's location)—the prohibition on violation of *jus cogens* norms is absolute (Criddle and Fox-Decent, 2009; Fox-Decent and Criddle, 2012).

2.3.2 The Strengths of Fiduciary Theory

This section will outline the strengths of Fiduciary Theory, which is an important task in order to justify precisely why it was adopted in this research. It will be argued that it is stronger than Social Contract Theory in describing the relationship between state and subject as it replaces consent with trust; that it provides a bulwark against the permanent state of emergency, and that like IE, it is a framework that can accommodate pluralism.

2.3.2.1 Trust, not Consent

In competing theory describing the relationship and arising duties and obligations between the state and subject (Social Contract Theory), an essential component building the foundation of this relationship and its correlative duties is that of consent. The subject consents to the sovereign's irresistible power. Fiduciary Theory replaces *consent* with *trust*.

Consent is a "fiction" and Fiduciary Theory builds upon this vacuum, positing "...a concrete normative structure that aspires to make rightful the possession and exercise of explicitly nonconsensual sovereign power," and "...insisting that every person must have an equal opportunity to participate in political processes that ultimately culminate in the state's possession and exercise of nonconsenual coercive power" (Cassinelli, 1959; Fox-Decent and Criddle, 2010, p. 315).

Legal authority does not flow from consent, it "...flows from the rule of law and goes to the authority of the state to announce and enforce law, to establish legal order rather than some other kind of order" (Fox-Decent, 2011, p. 89). The state's political authority grants it power to determine the "substantive" content of the law within constitutional boundaries, and its political authority is entrusted to it by law (Fox-Decent, 2011, p. 89). Fox-Decent (2011, p. 89) argues that "[t]ogether, legal authority and democratic political authority express the ideal of popular sovereignty, the notion that all public power

derives from the people." Legal authority demarcates the use of authority and mere power (Fox-Decent, 2011, p. 92).

The state is entrusted by law to establish legal order on behalf of its subjects—it is granted public power to be exercised to the subject's benefit, which is exercised on the basis of their trust. The subject trusts the state to discharge its fiduciary obligations, as failure to do so threatens its legitimacy and makes it accountable to the subject (Fox-Decent, 2011). As it turns out, the individual subject that may actively distrust the state and reject its "claim to authority", however must trust the state to act to its benefit (Fox-Decent, 2011, pp. 105–106).

The Fiduciary Theory accepts a factual relationship of asymmetrical power between those who govern and those who are governed—democracy is the ideal but the fiduciary principle is triggered independently of how power was obtained, it regulates the usage of power regardless of the fact (Fox-Decent, 2011).[35] Consent is a fiction that undermines Social Contract Theory as a viable alternative for explaining sovereignty.

It may be controversial to move away from consent as a justificatory principle on which state authority (and the correlative obligations of its subjects) is founded, though it might be noted that the concept of consent in such relations is in itself controversial, inspiring much debate and even seeing some of its proponents admit its weakness whilst trying to defend it (Cassinelli, 1959; Pitkin, 1966; Beran, 1977; Tuckness, 2016). There are those who argue that consent is essentially explicit through democratic participation in elections, that it is implicit in acceptance of membership of an association (the state), or that it is implicit where one simply *should* consent because of the *good character* of any given government (consent is tacit or hypothetical rather than explicit)—correlative duties to obey the state arise from acceptance of membership of the State and its authority, it is to some degree promissory (Pitkin, 1966; Beran, 1977). To an extent, some theorists conflate acquiescence with consent, however due to the state's coercive power there may be no option but acquiescence (Cassinelli, 1959; Fox-Decent, 2011). For there to be consent, there must be choice—if one is coerced into accepting authority, such as through the enforcement of law or fear of social sanction, one is not truly consenting (Cassinelli, 1959). That a subject should consent to a state

[35] Democratic process and liberal democracy are evidently the preferred legal and political model supported by Fiduciary Theory, however wielding discretionary power over a subjugated population still triggers the fiduciary principle—therefore even a government exercising de facto sovereignty is still required to wield its power on behalf of its subjects (Fox-Decent, 2011, p. 104).

based on its character or moral value is insufficient, it does not establish consent (it is hypothetical, and truly a fiction unless proven otherwise)(Fox-Decent, 2011, p. 141). Consent is represented more explicitly in elections to some degree, where the public votes for particular parties and their policy positions and the winning party has a popular mandate—even in this case however a popular mandate, while lending political legitimacy to the state as an expression of popular will, represents consent of the majority (insofar as the winning party adheres to its mandate), not of all subjects (Cassinelli, 1959; Fox-Decent, 2011).

Some would argue that there is choice and therefore consent, if at least implicit, as one has options to opt-out of the state—either through public declaration, migration, or secession, for example (Fox-Decent, 2011). Migration is perhaps the weakest argument here, as this is of course contingent on resources—even consent proponent Beran (1977) acknowledges that this is unworkable (Fox-Decent, 2011). As to secession, this too requires resources, as well as perhaps a sizeable group of dissenters sharing the same ideology; it is not a particularly accessible nor always entirely plausible option where the number of dissenters is especially negligible (Fox-Decent, 2011). As to the idea of public declaration and opting-out individually, at this point the sovereign citizen becomes judge and party in any conflict with the state's legal subjects, s/he exercises arbitrary power and becomes a threat to the secure and equal freedom of those in the state's territory, whose interests the state has been entrusted with protecting—the state must take responsibility for this 'sovereign' citizen whether they consent to it or not (Fox-Decent, 2011).

The contract approach is flawed, as contract like relations imply consent which is not universally present in the state subject relationship, and it is therefore not a strong conceptual resource to rely on in describing and prescribing such relations (Fox-Decent, 2011). The trust relationship is more accurate, as in a trust the trustee is granted power over the beneficiary's interests, to which no consent is actually required (Fox-Decent, 2011). It accurately describes the asymmetrical relationship between state and subject, recognising that the subject is vulnerable to the state's unilateral power but that this power is limited to establishing a regime of secure and equal freedom in the interest of the subject (it is the reason for which the fiduciary principle has legally authorised the state's assumption of power) (Fox-Decent, 2011). In return for this provision of secure and equal freedom under the rule of law, the state demands the subject's obedience

insofar as the laws which they construct are indeed law, not merely decisions or rules that act upon the subject in a manner that they could not conform with or anticipate, and so long as they are not contrary to their fundamental interests (such as human rights) (Fox-Decent, 2011).

2.3.2.2 State Prerogative and Constraints

The question of the state's prerogative and its exercise of power under exigent circumstances is a hotly debated and complex one. From a rather early point, John Locke argued that prerogative, the discretionary power to act without prescription of or against the law, was justified to adapt to exigent circumstances (Gross and Ni Aolain, 2006, pp. 118–120). It was to be adopted for the public good. Political realist approaches such as that of Carl Schmitt go even further. Schmitt (2005) argued that a state had to have unlimited power in the midst of exigent circumstances, and not only that but its decision of the "exception" proved its sovereignty. Schmitt (2005) believed that actions within the exception could not be reconciled with the law of normal time and that state action could not be constrained by the law. Schmitt (2005) believed that the exception subsumed the norm.

The subsumption of the norm by the exception is tantamount to permanent emergency, which Agamben (2005, p. 2) considered the "...dominant paradigm of government in contemporary politics." Emergency rule becomes the norm, or at the very least, can be difficult to end as Oren Gross and Fionnuala Ní Aoláin (2006, p.175) argue, "[e]mergency regimes tend to perpetuate themselves, regardless of the intentions of those who originally invoked them. Once brought to life, they are not so easily terminable."

The law is vulnerable to being warped by extreme circumstances, extreme measures can percolate into and "contaminate" ordinary law (Gross and Ní Aoláin, 2006, p. 161). The exception can also become normalised (Gross and Ní Aoláin, 2006, p. 228). [36]

To avoid the corruption of ordinary law, Gross and Ní Aoláin (2006; 2008) argue for a model of Extra-Legal Measures—that is, they condone a model of official emergency response where the state can act outside of the law, so long as responsible state officials

[36] Additionally, according to Gross and Ní Aoláin (2006, p. 228):

> ...as our understanding of normalcy shifts and expands to include measures, powers, and authorities that had previously been considered special, exceptional and extraordinary yet necessary to deal with emergency, the boundaries of new exceptions are pushed further to include new and more expansive powers and authorities.

present themselves to the public for censure or sanction. This approach is closer to that advocated by Fiduciary Theory but does not quite go far enough. While Fiduciary Theory holds that the state loses legitimacy and is accountable to its public when it fails in its fiduciary obligations, it does not follow that if the public indemnifies the state's actions the state committed no wrong. Because the people decide that an offence was either not an offence or should be forgiven does not override the commission of what could very well be an egregious act. Fiduciary Theory is infused with morality as well as legality—instruments of the government are required to discharge fiduciary duties independently of whether or not deviations will be accepted by the people.[37]

The question is one of the appropriate agency of the state, or to quote Nomi Claire Lazar (2008, p. 166), "[i]f we embrace agency, how do we constrain excess?"

Fiduciary Theory can help resolve the problem of the permanent, unregulated emergency. Its approach to emergency is arguably what Gross and Ní Aoláin (2006, p.35-66) would call constitutional accommodation—that is, emergency provisions are accommodated by law. Through its general endorsement of human rights, and ultimately (largely) the international treaties and practice that they entail as well as the inherent requirement of proper procedure being followed that is compatible with the requirements and jurisprudence of international law.[38]

Even if one argues that the ultimate source of accommodation is not found within international practice itself, that the state is ultimately self-regulating,[39] the procedures governing entry and conduct during crisis is *constitutive* of the fiduciary's obligations.

Fiduciary Theory rebuts Schmitt, the state's authority is circumscribed by its fiduciary obligations, by its required adherence to human rights. Fiduciary Theory requires the state to neither instrumentalise nor dominate its subjects—arbitrary application of power is prohibited. It is accepted that the state has a prerogative in how it responds to

[37] And regardless, if the public accepts an illegal act, contrary to fiduciary duty, the act commissioned is no less illegal.

[38] For instance, Gross and Ní Aoláin (2006, p. 256) argue that:

> ...certain international human rights treaties, specifically the European Convention, through the process of domestic incorporation as well as the garnering of "constitutional-like" status, through its influence on domestic judicial thinking, should be categorized as a form of constitutional rather than legislative accommodation.

[39] Which is in-keeping with the non-positivist tenor of Fiduciary Theory.

emergency, rights may be derogated from but under strict conditions.[40] As has been demonstrated, the state seeking to derogate in order to enhance its power is required to provide justification, measures must be necessary and proportionate, international notification and opportunity for contestation must be provided. The violation of peremptory rights is strictly forbidden. The derogations apply strictly for as long as necessary. Strict adherence to the fiduciary requirements preclude permanent emergency. This is normative theory that requires adherence by states and places a partial responsibility of ensuring compliance on the public and international community—the possibility of permanent emergency looms large still where States fail to comply with their obligations, and where publics and international actors fail to hold them accountable. However a wrong remains a wrong, whether or not relevant actors hold an offending State accountable, or indemnify a wrong, this normative framework remains a useful tool for prescribing action and describing deviations from obligations.

2.3.2.3 Pluralism

The underlying Kantian background to Fiduciary Theory may not strictly lend itself to pluralistic applicability, however at its most minimal the theory asserts that based on their legal relationship of asymmetrical power, states are bound to provide legal regimes of secure and equal freedom, requiring non-instrumentalisation and non-domination. Such requirements are preconditions for human dignity to flourish, which it should be uncontroversial to say is a universal concept. That human rights are constitutive of the fiduciary relationship also lends the theory universal appeal and applicability.

Jack Donnelly (2013, pp. 94–96) argues for the universality of human rights (whilst recognising their "particularities") on three grounds:

- Almost all states consider internationally accepted human rights to be an entrenched aspect of politics and law.[41]
- A plurality (almost all) of cultures, religions and worldviews participate in an "overlapping consensus" on internationally recognised human rights.[42]

[40] Deviations from law then must in themselves be lawful.

[41] Donnelly (2013, p. 94) states that "...the six core international human rights treaties—the two Covenants [*International Covenant on Economic, Social and Cultural Rights,* and *International Covenant on Civil and Political Rights*] plus the conventions on racial discrimination, women's rights, torture and the rights of the child—in early 2012 had, on average, 172 parties."

[42] Donnelly (2013, p. 96) argues that the moral equality of humans is endorsed by "...most leading comprehensive doctrines in all regions of the world." It is a convergence between and within

- This consensus is based on a universal recognition of contemporary threats to human dignity.[43]

Human rights have reached a wide level of acceptance, in terms of their utility and importance to modern international relations. The concept of human dignity has united a plurality of views around the premise of human rights, the importance of which concept is recognised by Fiduciary Theory, and the importance of human rights as being constitutive of a state's rule should lend it some degree of cross-cultural palatability.

2.4 The Rationale of the Dual Theoretical Framework

The decision to use a dual or bi-partite theoretical framework is an ambitious one but one which is necessary and holds practical value.

This research is concerned with a complex set of interacting factors; disaster, emergency, the design of an ethical EMIS, and the ethical deployment of such systems by state agencies that respects human rights.

The challenge ahead may be best approached from two streams of analysis, as each one, on its own, may fail to satisfactorily engage with the full range of issues.

An ethical approach is taken generally with a view towards assisting in the design of a system that performs morally above and beyond the requirements of law, which can stagger behind technological development. The specific context of the system is one which is used to save life and property in disaster. Such a system will have two ethical goals, the preservation and management of the environment, and the preservation of human life and dignity. This involves a complex network of agents. The informational ontology used by IE can enable a holistic and thorough identification and examination of the agents and patients involved in this network, and the deduction of how a) these agents can perform ethically while causing minimal harm to patients and b) how these agents can perform together to pass through the moral threshold, or how distributed morality can occur. This approach is inclusive; it will examine how a system can be designed to aid an ethical response to threats to the environment and humans.

---

civilizations, Donnelly (2013, p. 96) argues that "...provides the foundation for a convergence on the rights of the Universal Declaration."

[43] This argument Donnelly (2013, p. 96) calls functional universality, the basis of which is that human rights are the most effective protection against contemporary, common, and global threats to human dignity posed by states and market economies.

The second stream of analysis will be more anthropocentric; it will be concerned more specifically with how state actors deploy software systems in emergency management and will examine the constraints that they face. Fiduciary Theory is useful in this instance, as it provides normative guidance on the duties of the state and its constraints, it can help deduce how they should observe human rights—arising from which are both positive and negative obligations—in extreme circumstances.

The dual analysis presents some opportunity for convergence. Both frameworks, though operating at different LoAs—the wider informational environment and the relationship between state and subject—are compatible. Both are responsive to the concept of human agency and how it is intricately linked to dignity; they essentially share a common value of flourishing (human and in the case of IE, ecological). By initiating an analysis of the interactions between agents and patients in the infosphere using the ontology of IE it should be possible to transplant some of these moral evaluations into the moral foundation of Fiduciary Theory. IE is used, ultimately with its novel ontology, to identify new threats to human dignity and add substance to the fiduciary evaluation. IE asks what is good in the infosphere and what is good for the infosphere? Fiduciary Theory outlines the obligations of states, grounded in the factual, legal relationship between state and subject. By examining the ethical substance of issues that affect human beings (but in which humans are not the only agents), it should be possible to re-evaluate the state's fiduciary obligations and offer a more refined, persuasive argument on the fiduciary requirements of states in a post-Westphalian and globalised world.

Finally, both frameworks have a pluralistic appeal. This is important in a globalised world where actions in one region often reverberate internationally. It is important to endeavour to make ethical evaluations that can be acceptable throughout culture and place. This is ambitious, and claiming to come to 'culture-proof' evaluations would be absurd—however the researcher has endeavoured to use frameworks that support shared values and common solutions, and presents this work as a contribution to the discussion of ethics in an inter-connected world. The conclusions found here are open to disagreement, which are welcome. Only by having the discussion can we eventually come to mutually agreeable solutions. And these discussions must happen now, as the world evolves around us and the nature of our ethical challenges change, we cannot be complacent and must be vigilant in the regulation of our new technologies.

2.5 Conclusion

This chapter has offered succinct explanations of both IE and Fiduciary Theory and argued in their favour against alternatives. This research will examine complex interactions between human and non-human agents and is concerned with how these interactions can harness distributed morality without harming the moral patients of the processes embedded in the moral situation. IE provides a useful ontology that is accommodative of value pluralism.

Secondly, the research is concerned with constraints on state actors in emergency that use EMIS. Fiduciary Theory offers normative guidance on the obligations of states, particularly in emergencies, and the constraints on their actions. It too has pluralistic value. Information ethics can dialogue with or inform Fiduciary Theory to better and more persuasively frame the nature of the state's obligations to its subjects in the infosphere.

3 METHODOLOGY

3.1 Introduction

This chapter will explore the methodological approach taken in this research, and seek to justify the data collection methods utilised as well as explain how this data can be analysed fruitfully using the theoretical frameworks described in the previous chapter.

In this chapter, it will be argued that the most appropriate theoretical model in the context of this research is a constructionism informed by the Philosophy of Information (PI). Reality is interpreted by epistemic agents (humans), who give data meaning and shape the world around them with information. The approach is useful for opening a set of data collection tools, crucially including interviews. If reality is constructed by people, then constructing reality is a collaborative venture that invites social interaction. It will be argued the PI approach will be particularly useful in investigating systems in a theoretical or propositional form.

The form the research will take will then be outlined, which is a disclosive analysis. The disclosive analysis is a value based appraisal of a described technology—it seeks to bring transparency to the unknown workings of a technological system. It will be argued that this is a useful approach in uncovering the morality in the design and use of computer systems and anticipating a system's impact on moral values.

This chapter will conclude by providing an overview of the precise methods used for data collection which are in descending order of importance: semi-structured interviews, observation, and document analysis.

3.2 The Methodological Approach

3.2.1 The Qualitative Approach

From an early stage the methodology chosen for this research was qualitative. The questions that this research seeks to ask and address are not ones which are better investigated or answered using quantitative methods. The research is concerned with the relationship between an information system that is (or will prospectively be) deployed in emergencies, and moral values. Whilst generally quantitative methodology, or perhaps even a mixed methodological approach, might yield useful results in this research area, the particular scope of this research limits the tools that can be used

effectively in answering the research questions posed by this project. After consideration, the qualitative methodology proved to be a superior approach for the execution of this research, providing more effective tools for data collection and analysis. At the heart of this research are philosophical and legal questions. Fundamentally, the questions that are asked are "how" and not "how many?", and as such a qualitative approach is more appropriate and fruitful (Silverman, 2013, p. 12).

The research is concerned with modelling a system under development and examining its potential ethical and legal ramifications vis-à-vis particular societal values at a pre-deployment stage—the particular context of this research does not overtly support quantitative exploration.

In order to collect relevant data that can be analysed with the selected theoretical frameworks, interviews with relevant experts on the Slándáil project were selected in order to deduce the capabilities and uses of the in-development technology under study.

3.2.2 An Informational Epistemic Orientation

Following the decision on the methodological approach, the next challenge was deciding on an epistemic orientation that would best frame and guide the research (Silverman, 2013, p. 105).

The epistemic orientation, "...provide[s] an overall framework for viewing reality" (Silverman, 2013, p. 105). Therefore, selecting an appropriate epistemic orientation is integral to conducting coherent research, in guiding decisions in research methods and subsequent analysis that facilitates cogent results consistent with the frame of reality that was provided.

The most appropriate epistemic orientation would be one which unlocks research methods that, as indicated, allow the modelling of a system of objects and not only analysing the interactions between those objects as they exist, but analysing their potential relations under theoretical conditions. This research is concerned with the impacts of a particular system type, essentially a hybrid human and technological multi-agent system that has yet to be implemented in a manner where its impacts can be observed and recorded. It requires a fluid epistemic orientation through which it can be examined. In the context of emerging technology with ethically loaded value, to wait until one can record and observe it active in a 'real life' situation would simply be

irresponsible when a timely and practical ethical and legal analysis should be able to anticipate any ethically problematic system behaviour ahead of its implementation, before it can cause harm and so that potentially harmful aspects of the system can be addressed before it is indeed implemented.

Fortunately, from an early stage in the life cycle of this research, it was decided that Information Ethics would form a component of the dual theoretical framework, and whilst it may seem counterintuitive to begin with a theoretical framework before a theoretical model, the Philosophy of Information (the branch of philosophy from which Information Ethics was born) indicates the most useful epistemic orientation that overcomes the relative shortcomings of other approaches (Greco *et al.*, 2005; Floridi, 2008, 2011a).

The epistemic orientation taken might be said to be a counterintuitive one that involves something of a marriage between constructionism and a particular kind of realism (in this case, information structural realism—or ISR). Floridi's (2011a, 2011b) approach to epistemology and solving philosophical problems generally involves a multi-component process of *minimalism*, the *method of abstraction* (with which the reader should be familiar from Chapter 2) and *constructionism*.

Floridi (2011a, p. 285) provides a thorough defence of constructionism, arguing that "...knowledge neither describes nor prescribes how the world is but inscribes it with semantic artefacts". For Floridi (2011a, p. 291) "...knowledge is acquired through the creation of the right sort of semantic artefacts, information modelling, in other words," and "[w]e are the builders of the infosphere we inhabit...". In this mode of knowledge acquisition and creation, Floridi (2011a, pp. 292-293) argues that, "...experiments do not imitate the world, they shape it". The human mind is necessary to bring meaning to the universe, it constructs reality—interpreting data around it and constructing information. What is real and what is knowledge is thus because it has been assigned meaning. According to Floridi (2011a, p. 291), in knowledge, and knowing reality, "...knowledge becomes a collaborative enterprise of growth and refinements in a multi-agent system (humanity)."

Floridi (2011a, p. 293) suggests that constructionism, combined with minimalism and LoAs, can be used to answer questions "...that are not answerable in principle empirically or mathematically...", which for this research poses an attractive avenue to

explore—the research is propositional in many ways, and the tools that constructionism provides when combined with others (minimalism and LoAs) lead to the conclusion that constructionism may indeed be a suitable approach. This warrants further unpacking, as well as a brief explanation and justification of ISR. A step back will be taken from construction in order to unpack the concepts and uses of minimalism and LoAs in this research.

Prior to constructionism in the process of solving philosophical problems in research are minimalism and LoAs. It is argued that discrete systems can be chosen to improve the "tractability" of the problem space (Greco *et al.*, 2005, p. 624). Minimalism directs the choice of the philosophical problem using three criteria: controllability, implementability and predictability. The problem space in this case was pre-selected, however the criteria justify the initial decision.

Gian Maria Greco *et al.* (2005, pp. 624-625) provide an explanation for these three criteria. On controllability and its use: "[a] system is controllable when its structure can be modified purposefully. Given this flexibility, the system can be used as a case study to test different solutions for the problem space" (Greco *et al.*, 2005, p. 624).

On implementability and its use (Greco *et al.*, 2005, p. 624):

> The second minimalist criterion recommends that systems be implementable physically or by simulation. The system becomes a white box [open and knowable to those who construct it–in this case, the researcher], the opposite of a black box. Metaphorically, the maker of the system is a Platonic "demiurge", fully cognisant of the components of the system and of its state transition rules. The system can therefore be used as a laboratory to test specific constraints on the problem space.

Finally, on predictability and its use: "...the chosen system must be such that its behaviour should be predictable, at least in principle. The demiurge can predict the behaviour of the system in that she can infer the correct consequences from her explanation of the system" (Greco *et al.*, 2005, p. 625).

In the case of this research, the problem space is the use (and ethical/human rights implications thereof) of social media harvesting emergency management information systems in response to natural hazards. The system chosen for analysis is the Slándáil EMIS and the agents/patients that comprise and are affected by it, which will be addressed in more detail presently. The study of this system fulfils the three listed criteria. The system is controllable as its structure can be modified purposefully using

the method of abstraction, which will be revisited imminently. It is implementable; through acquisition of information and construction of a descriptive account of the system and its functionality it becomes a white box. The system is predictable to a degree—upon construction of the white box the researcher can anticipate to some extent its potential impact on the problem space.

Next, the reader may recall the previous explanation offered of the Method of Abstraction: "A level of abstraction is a collection of observables analysed with a particular goal, where everything but the observables relevant to the analytical goal are abstracted." This method is used to analyse "discrete systems", applying to "conceptual" and "physical" problems (Greco *et al.*, 2005, p. 627).

Upon rendering a LoA, or multiple LoAs featuring different observables, the system can be simulated—the models produced can be logically tested with different variables. Simply put, "...a simulation is considered the observation of a model that evolves over time" (Greco *et al.*, 2005, p. 627). In the course of this research, each level of abstraction is intended to encapsulate different objects in the emergency management scenario, including the software artefacts that comprise the EMIS under study, and the human agents on either end of that system (emergency managers, technologists, social media users and disaster survivors). In essence, and to borrow terminology again from Floridi, LoAs should map relata and their relations (the nature of the network of objects that comprise a system)—it is important to understand characteristics and properties of these objects and the behaviours of the system under study (Greco *et al.*, 2005; Floridi, 2008, 2011a).

Reality is therefore modelled through LoAs, which warrants a return to discussion of ISR. The very concept and purpose of LoAs may remind the reader of the realist epistemic orientation, due to its structural concerns and concern with describing systems and their causal powers. ISR supports an informational view of reality, and arguably forms the ontological foundation of PI and IE. According to Floridi (2008, p. 236), ISR is "...committed to the existence of a mind-independent reality addressed by and constraining our knowledge." Under ISR, the objects which comprise systems are mind-independent, informational objects—"...cohering clusters of data...", or "...concrete points of lack of uniformity" (Floridi, 2008, p. 236). Informationally then, something simply is or is not. This may initially seem to contradict the constructionist approach, but it arguably complements it. Data simply exists, and is there to be discovered, and when

it is, it follows that it has to be modelled and given meaning to truly constitute information—data is found and semanticised, socially constructed into information. Informational structures require epistemic agents (humans) to decode (Floridi, 2008, p. 247).

An advantage for the researcher of ISR and LoA is that due to the informational nature of objects, they need not be physical or even exist beyond a propositional form. These objects that can be added to the LoA can be propositional or hypothetical: a system being investigated "may be entirely abstract or fictional" (Floridi, 2008, p. 226). This approach serves the current research well, it is investigating an unactuated system, and will require the consideration of objects that may only exist essentially in propositional form.

The PI approach subscribes to the philosophical epistemic tradition of the maker's knowledge—that "...one can only know what one makes...", and therefore the researcher, as a Platonic demiurge, must obtain data and construct a white box—a system where the internal structures, rules and compositions are known and disclosed (Greco *et al.*, 2005, p. 629). The researcher becomes the creator, constructing a reality that is known by themselves and in the context of this research, transparent and known to the reader.

Constructionism then is an approach that is compatible with the goals of this research. Reality is pieced together by the researcher, who investigates the general composition and behaviours of systems and the relations between constituting objects which do not need to be empirically perceivable. This is a fluid and dynamic epistemic approach that unlocks numerous methods. Observation and interviews are viable methods of data collection that can help the researcher construct reality. Because reality is socially constructed, even if data vis-à-vis reality can exist independently of the mind but requires decoding by the epistemic agent, reality building can be collaborative, where people (researcher and interview participants) can produce it together with their shared insights.

3.3 Disclosive Computer Ethics

The form this research will take is fundamentally a case study. The subject of the case study is a social media powered EMIS (Slándáil). At a more complex level however the purpose of the case is a disclosive analysis which will analyse the impact that this system

has upon societal values—modified to also consider its implications for not just values from an exclusively ethical perspective, but to examine tensions between the system and human rights. Fiduciary Theory and Information Ethics will anchor this analysis.

In the following, time will be taken to briefly reacquaint the reader with Slándáil, and then to justify the value disclosive analysis approach taken.

3.3.1 Slándáil

The reader should recall Slándáil from Chapter 1.[44] The system has not yet been deployed in the field, and as such, until it is finalised and becomes a *bona fide* element of emergency response, actively utilised by emergency management professionals, it remains more of a propositional entity than an 'actual' one. Though limiting opportunity for data collection, investigating an in-development system should not be viewed as problematic, but as an opportunity. Investigating this system allows for a thorough analysis of the potential harms that similar systems can perpetuate before they have been deployed—acting expediently in ethical/legal analysis before they become a mainstream aspect of emergency management enables the identification of possible issues that may require addressing.

Slándáil will be investigated by the researcher in order to deduce the implications for moral/societal values, including life, privacy, justice, trust, accountability and responsibility. These values were chosen after an early literature review based on perceived importance, weighed against the time and space available to the researcher. These values are often implicated with the emergence of new ICTs, and whilst others are too (transparency and autonomy, for example, come to mind), the previously listed values were decided to require analysis with particular urgency.

[44] To remind the reader, it was described as thus:

> Slándáil is an EU FP7 funded project lead by Trinity College Dublin in collaboration with partners across Europe including academic, business and emergency response actors. Representing the convergence between conventional emergency management information systems and technology that harvests structured information from social media, the project seeks to establish a system that ethically harvests relevant information from social media during emergency response to natural hazards (such as floods) that can contribute to situational awareness and provide decision support for emergency responders. The system will be a combination of emergency management software and text/image analytic software.

3.3.2 Disclosive Ethics and Slándáil

The purpose of the disclosive analysis in (computer) ethics is to bring clarity to the opacity of technological systems, in order to understand the moral properties embedded in technological systems and therefore their ethical impact (Brey, 2000, 2010).[45]

Disclosive ethics are compatible with the constructionism discussed here, in building a white box of the investigated technology, the technology and its applications become transparent—both to the researcher and the reader. The disclosive approach is concerned with the description of a technology, and the deduction of its design and applications on moral values (Brey, 2000, 2010). It is essentially concerned with making the unknown known, so that its "hidden morality" can be exposed and analysed (Brey, 2000, p. 126). The disclosive approach is a carefully descriptive exercise then, entailing a thorough description of the investigated technology and its moral import (Brey, 2000, p. 127).

Brey (2000, 2010) suggests that the investigated technology be analysed in something of a thematic manner, with a focus on how the investigated technology impacts predetermined societal values. Brey (2010, p. 53) provides his own list of suggested values including "...justice (fairness, non-discrimination), freedom (of speech, of assembly), autonomy, privacy and democracy", and adds that "[m]any other values can be added, like trust, community, human dignity and moral accountability."

The disclosive study of technology is a two-stage process. Of the first stage, Brey (2000, p. 127) says:

> In the first stage of analysis, some technology (X) is analyzed from the point of view of a relevant moral value (Y)(where Y is, e.g., privacy, justice, freedom, etc.), which is only given a loose, common sense definition. This analysis may yield a tentative conclusion that certain features of X tend to undermine (or perhaps sustain) Y in particular ways.

[45] Philip Brey (2000, p. 126) provides an instructive rationale for the disclosive approach:

> ...I want to claim that a large part of work in computer ethics is not about the clarification of practices that have already generated moral controversy, but rather *revealing the moral import of practices that appear to be morally neutral*. Many designs and uses of computer systems, I want to claim, have important moral properties, that remain hidden because the technology and its relation to the context of use are too complex or insufficiently well known.

On the second stage, moral theory is utilised through application to the context of the technology-value relation and potentially developed further based on the analysis and its particular requirements, should existing theory not adequately be equipped to satisfactorily address the normative aspects of the situation (Brey, 2000, p. 127). Brey (2000, p. 127) expresses scepticism about a more theory-driven approach to the disclosive analysis, as "...a theory-driven approach tends to make the acceptance of a disclosive analysis dependent on the acceptance of a particular moral theory." A loosely defined approach to the problem area (the value in question) can be more persuasive than more specific, theory-driven ones. Brey (2000, p. 127) also argues that theory-driven approaches may have an inherent bias distorting analysis, including "preconceptions" about the technology being investigated. The disclosive analysis conducted in this research, however, will be theory driven and will include IE and Fiduciary Theory as the foundation of the theoretical analysis.

On the contrary to Brey's arguments, a loose definition of values fails to adequately explain their substance, and privileges palatability over analytical depth. One can agree why a value is just that, and the purpose it serves in society at a *prima facie* level (like privacy, for example), but without an adequate investigation of its substance and arising normativity, analysis is doomed to be thin and nebulous. The core of the value must be unpacked, so that the precise implications of the investigated technology can be analysed in a structured way against particular principles, lest the analysis risk being vague, opaque and ultimately less persuasive. Rather than opting for general approaches to values, this research will use the dual theoretical framework which, whilst it has principles built in, is engageable from multiple philosophical traditions—it has pluralistic appeal without being reductionist to the point of operating on a level of intuition. The risk of contention is favourable to operating on a generic level where intuition supersedes a more precise analytical framework.

Brey's (2010) second issue with a theory-driven disclosive analysis, that it may embed presuppositions, is more a problem caused by the researcher than the theory that they consult, therefore should not be considered an issue—theory should not be used if it is already inherently biased against certain practices without at least first enabling logical discourse on why something may or may not be wrong.

Brey (2010) advocates an inter-disciplinary approach to the disclosive analysis, one which requires the talents of philosophers, legal experts, social scientists and

technologists. The current researcher's academic background positions him well to occupy the first three roles within reasonable limits, and in conducting the research was able to draw on the knowledge and expertise of technologists, who were essential in explaining the functionality and capabilities of the technology under investigation.

This research then will approach analysis in the manner prescribed by disclosive computer ethics, with a focus on the theory outlined in Chapter 2. In the following chapter, the Slándáil EMIS will be described, that is, its components and proposed uses will be laid out and made transparent. In the subsequent chapters, the implications of the system's design and applications will be analysed by its impact on values such as privacy, justice, trust, responsibility and accountability—using the dual framework to offer normative guidance and substance to the analysis.

3.3.3 Disclosive Analysis as a Case Study: The Issue of Generalisability

This disclosive analysis, which will require the collection of data on a particular case and an analysis of this case means that the research will be using case study method.[46] The case study will investigate the behaviour and impact of "...a set of actors engaged in a sequence of activities... over a restricted period of time..." (Mitchell, 2006, p. 169). The set of actors will be software artefacts, technologists and emergency management professionals. The goal will be to understand the set of actors as they exist as a system and investigate the impact of this system on values. The case will need to be investigated in order to construct the correct LoAs that allow fruitful analysis.

The system, the Slándáil EMIS, is intended to be representative of EMIS that process data from social media into actionable information for emergency managers—it is a pathway to exploring what such technologies can be capable of. The system, to use the language of PI and IE, is a token and not the whole type—this is a single case study but one which is conducted with the intention of producing general conclusions that can apply to the type and not just the token (Floridi, 2013). Such an endeavour, as with the single case study, might raise questions of the generalisability (Silverman, 2013; Gomm, Hammersley and Foster, 2006). A valid question to raise is the applicability of conclusions derived from the single case to the wider context, or how representative the single case can be to the wider context of similar cases.

[46] A case study, "[i]n its most basic form... refer[s] to the fundamental descriptive material an observer has assembled by whatever means available about some particular phenomenon or set of events"(Mitchell, 2006, p. 168).

The primary research question here is "what are the societal value/human rights implications of Slándáil-type systems and how can value threats be mitigated?". The case study is undertaken with a view to assessing the capabilities of a specific technology with a view to discerning possible value impacts, but such features can and likely would be replicated across any other technologies that perform the same functions towards similar goals by varying degrees. The case study may be a token but will be a portal to the type.

There is no panacea or magic bullet for generalising on the single case, therefore it is necessary to refer back to the previous section. Recall that LoAs could contain fictitious objects—in using the Slándáil platform as a case study, the objects constituting the entire system will not be the only ones observed, fictitious aspects can and will be included, the researcher will hypothesise the addition of software artefacts and additional variables that do not constitute the system but may viably constitute it in order to more thoroughly assess its capabilities and capacities for harm (or help). A gradient of abstraction can be formed that can examine the system under different, hypothetical, conditions and the case will not use only actual observables (Floridi, 2008). With a gradient of abstractions including propositional objects, the generalisability will be increased by enabling the researcher to analyse the moral situation with additional, propositional elements. Limits remain, the researcher remains restricted by the information to which he has access in constructing a reality that can be generalised. The applicability of conclusions derived from investigation of the case should be high where other cases are similar, however it should be informative in guiding the decisions of those who design and implement similar systems and it can be a useful academic resource to those who wish to evaluate similar systems.

3.4 The Methods of Data Collection

3.4.1 Semi-Structured Interviews

Semi-structured interviews were selected as a primary method of data collection. Sampling was purposive (Silverman, 2013), and interview participants were selected based on their expertise and the nature of their contribution to the Slándáil project. Five technologists were selected based on their involvement with integral aspects of the system's functionality including aspects such as text processing, data aggregation and analysis, geo-spatial systems, image analytics and programming. Technologists were based in Dublin (Trinity College Dublin), Germany (CID), and Italy (DataPiano). Two

emergency management professionals were interviewed based on their expertise and knowledge in emergency management and the context in which the system would be deployed; they were from An Garda Siochana, and Police Service Northern Ireland (PSNI). The choice of interview was selected as the researcher could solicit information directly from persons with a close knowledge of the system that was under-investigation—a maker's knowledge of a sort. The participants could carefully explain the features of their work, and researcher and interview subject could engage conversationally about the theoretical capabilities of the system. Therefore, by interviewing technologists the researcher could deduce the characteristics and properties of the investigated system and examine how propositional additions could alter it. Interviewing emergency management professionals invited knowledge from the users, who would be in the greatest position to explain the parameters of deployment of the system and how it would change the current emergency management landscape.

Interviews were conducted face-to-face, via Skype where geographical distance was an issue, and over phone and by email where brief follow-up was required. Each interview with one exception was recorded either by dictaphone, or using audio recording software when Skype was used. In the exception, one participant was only available by phone and notes had to be taken by hand.

Semi-structured interviews were selected in order for the researcher to control the pace and seek clarification on any ambiguities that may emerge as well as elaboration on useful, unexpected information.

Questions that could solicit useful data that could eventually aid the construction of LoAs were something which required consideration and there were numerous issues to consider. William Foddy (1994, p. 17) provides four steps that should be followed in a successful question and answer sequence, the four of which informed the structure and content of questions:

> (a) the researcher must be clear about the nature of the information required and encode a request for this information;
>
> (b) the respondent must decode this request in the way the researcher intends it to be decoded;
>
> (c) the respondent must encode an answer that contains the information the researcher has requested; and,

(d) the researcher must decode the answer as the respondent intended it to be decoded

These points establish the importance of being clear and forthright in an initial request from participants. A point a) failure will invariably cause a failure at points b) and c).

As such—to reduce risk of misunderstandings—prior to interviews participants were contacted by email and the purpose and types of information sought were made clear to candidate participants. Subsequently, informed consent forms were sent electronically that also provided clarification on the purpose of the interview. Foddy (1994, p. 71) cautions that when participants are under-informed about researchers' purposes, they form their own hypotheses that will influence their answers to questions, potentially to the detriment of their validity. The opposite problem, where a participant may be adequately informed but perceive the research as being against their interests may also pose a problem (Foddy, 1994, p. 72). The benefits of properly informing the participant of the research goals may outweigh the potential pitfalls, however, as Foddy (1994, p. 72) argues that "...*respondents who know why a question is being asked are in a better position to help a researcher than those who do not*: for example, by correcting any incorrect presuppositions that the researcher appears to have made." Due to the semi-formal nature of the semi-structured interviews, participants were given opportunity to engage bi-directionally with the researcher, allowing them as well as the researcher to seek clarifications.

A related problem stemming from willingness to co-operate based on information received by the participant from the researcher is that of question threat—where particular questions adversely affect a participant's willingness to co-operate due to a perceived punishment arising from answering that question (where questions might be embarrassing or harm their interests) (Foddy, 1994, p. 127). The researcher considered that due to the nature of the research, which investigates ethical and human rights issues, participants could plausibly feel uneasy with a line of questioning that was either combative or suggested that their own practices were unethical or may lead to unease with appearing unethical in their practices when the research was published.

Question threat was reduced by offering confidentiality of responses, as indicated as a solution by Foddy (1994, p. 112). No participant is referred to by name in this research. Questions were posed in a neutral manner, implying and assuming no deviance by participants. More sensitive questions designed more specifically to elicit responses

relating to potential areas of ethical concern in the participants' line of work were left to the end of interviews in order not to disrupt answers to less threatening questions. Questions also did not relate to the personal behaviour of the participants, but related to generic duties and properties and characteristics of systems on which they worked, thereby distancing the question from them on a personal level that might make them feel threatened.

A further point in reducing question threat should be outlined, which is that the researcher was able to "[e]stablish [a] lack of interviewer gullibility," as suggested by Foddy (1994, p. 125). The researcher has far reaching access to Slándáil project documents and staff and is known to be knowledgeable about the overall research of the project. To this end, participants could assume that any attempt at obfuscation would be ineffective. The researcher's position also placed him in a position of trust with interview participants, generating good will and pre-disposing them towards honesty and co-operation. A risk in this case is that due to assumed pre-existing knowledge, respondents would either give incomplete or overly complex answers (in terms of specialised language and explanation) (Foddy, 1994)—this risk was mitigated by encouraging participants to answer elaborately regardless of any perceived pre-existing knowledge of their work; during interview where answers were complex, the researcher engaged the participant with further questions until complex concepts and language were clarified.

Questions were constructed to be clear about the type of information sought. The use of double questions was minimised and no questions were leading (Foddy, 1994, p. 182). Vocabulary used was also appropriate and understandable to participants, who were experts in their respective fields—questions were simple and brief (Foddy, 1994, pp. 40–50). The researcher attempted to order questions in a descending level of complexity to the extent that question themes would allow, in order to prevent any adverse influences from earlier questions affecting answers supplied to later questions.

On point d), to ensure that questions were decoded properly by the researcher, interview participants were available for further questions in the event that clarification on answers was sought.

3.4.2 Observation

Observation was also utilised as a tool for data collection, however not strictly in the naturalist, ethnographic sense. The researcher was given the opportunity to attend numerous workshop and plenary meetings where project beneficiaries discussed their work, goals, aspirations and desires in terms of the technology under development and also (in the case of the technologists) demonstrated the software components of the system for which they were responsible. Attending these events enabled the researcher to familiarise himself with the technology under development and its potential utility for emergency managers. It provided a platform for taking notes that would direct appropriate lines of questioning during research interviews, where fuller explanations and clarifications could be sought.

In addition to this, the researcher was provided with access to various important aspects of the system, including the CID data analysis system (Topic Analyst), and DataPiano's SIGE EMIS. Experience with the software artefacts also served to direct an appropriate line of questioning, by inspiring questions about features witnessed and enabling a more than abstract understanding of the nature of the systems on which research participants worked.

3.4.3 Document Analysis

A library of research relating to the project was also available to the researcher. The library was extensive and encompassed more than research pertaining to the functionality of the system but also research relating more generally to how social media use could be maximised during emergency management. Documents were often in a technical language outside of the researcher's disciplinary field. Documents served to guide some the line of questioning during interviews, but were not useful as a primary source.

In one instance, a legal deliverable, *D2.6 Licence for the Use of a Disaster Management System* (Corbet *et al.*, 2017), was analysed thoroughly due to its salient implications for the governance of Slándáil-type systems. This analysis will be revisited in Chapter 8.

3.5 Research Ethics

The researcher has adhered to the standards of the *TCD Policy on Good Research Practice*, and *Ethical Guidelines of the Sociological Association of Ireland* (Sociological Association of Ireland, no date; Trinity College Dublin, 2002).

The researcher has carefully considered the ethical implications of interviewing participants for the purpose of this research, including potential privacy concerns or career impacts of the sharing of potentially classified information.

Interview participants were advised in consent forms (detailing their rights as interview participants) that the interview results will be available in a PhD thesis, that they were under no obligation to participate, and that their answers would be anonymised to the greatest extent possible. As such, participants will not be referred to by name in this research. All data was stored securely on a password protected laptop and cloud based service (Google Drive), with exclusive access to data by the researcher. Participants were given the option to retroactively decline consent before publication of thesis, at which point researcher could destroy data to the extent that this is possible.

One potential ethical concern may be that the researcher interviewed participants alongside whom he worked on the Slándáil project. The researcher has been transparent in sources of funding for this research. The researcher has conducted this research with integrity, and has striven to deliver an objective, accurate and transparent picture of reality through fair yet uncompromising questioning of interview participants, and critical analysis of the relevant issues. Proof of integrity is in the research that follows, which admits without any efforts at obfuscation that the technology under study does indeed have the *very real potential* to be used towards malevolent ends, and as such there are obvious risks and dangers posed to moral values and human rights. It is such plausible uses that necessitate this research, for it is only through understanding the evils that persons could implement such technologies towards, can one propose ways to prevent such uses. It must also be emphasised that the research is not intended as an evaluation of the outputs of the Slándáil project, but a more broad exploration of the value impacts of the technologies these outputs represent.

3.6 Personal Statement

This research is undertaken with the objective of analysing the relevant values and issues without bias, yet the researcher must acknowledge their own status and cultural and socio-economic background, factors which shape and influence their frame of reference in ethical and human rights analysis. The researcher is a White, Western European male from a working class upbringing, who was exposed prominently to Western European culture, attitudes, and value interpretations.

As stated earlier, it is important to endeavour to make ethical evaluations that can be acceptable throughout culture and place. This is ambitious, and claiming to come to 'culture-proof' evaluations would be absurd—however the researcher has endeavoured to use frameworks that support shared values and common solutions, and presents this work as a contribution to the discussion of ethics in an inter-connected world. The conclusions found here are open to disagreement, and refinement throughout time based on deliberation from diverse perspectives.

3.7 Conclusion

The foregoing has provided justification for the epistemic orientation providing a framework for reality in producing a case study and analysing it meaningfully. Constructionism will be used as PI argues it should; a white box (the case study) will be built by the researcher in co-operation with knowledgeable experts who will aid the researcher in piecing together reality. Data will be collected primarily through semi-structured interviews, allowing the researcher to control the pace of the interview and seek clarification and elaboration on issues as they arise.

The EMIS and its environment, and its networked agents, will be modelled at different levels of abstraction. The constituent objects of the white box will be actual and propositional.

The method of analysis will be a theory driven disclosive analysis, the white box will be held to a microscope and scrutinised based on its impact on values using the theoretical frameworks of Information Ethics and Fiduciary Theory. The white box that is built will assess these value impacts not only based on what the system is likely to be, but by inserting propositional objects not guaranteed to be integrated into a final system, also what it could be and in so doing, allow for greater generalisability and applicability of conclusions.

4 CASE PROFILE OF THE SLÁNDÁIL EMERGENCY MANAGEMENT INFORMATION SYSTEM AND INITIAL VALUE ANALYSIS: LIFE

4.1 Introduction

The purpose of this chapter is to begin exposing the hidden morality, or moral properties, of the Slándáil EMIS in order to understand the implications of such systems for the societal values of life, privacy, trust, justice and accountability.

This chapter will also initiate analysis on potentially the most beneficial aspects of the system, that is, its capacity to contribute to saving lives. In that respect, this chapter will hold the dual function of not only beginning to understand the Slándáil system but also the implication of such systems for perhaps the dearest of all values, life.

To locate the system within the theoretical framework of IE, the first task of this chapter will be to explain the concept of Distributed Morality, and the delegation of morally loaded tasks to artefacts. This is necessary to understand the potential of such systems to do moral good (or harm) when existing in a network of agents.

Following this, the actual and potential functionality of the system will be described in order to understand the functionality and utility of such systems broadly, and it will be argued that at a given level of abstraction such systems can qualify as agents based on Floridi's criteria of autonomy, interactivity and adaptability.

Upon achieving this, the concept of the tragedy of the Good Will shall be described and it will be argued that by delegating a morally loaded task to an artificial agent such as Slándáil, the power of Distributed Morality can be harnessed in order to save lives and therefore escape the tragedy of the Good Will.

Finally, a human rights analysis will be conducted on the relationship between such systems and the right to life in natural disaster situations, and it will be argued that whilst utilising such a system cannot be considered the totality of a state's duty in natural disaster response, it is a useful resource and potentially a state's very responsibility to adopt in emergency response if it is truly effective and feasible. It will be argued that its recording capabilities mean that it can assist in investigations into

decisions made during emergency response (a procedural aspect of the right to life) and also that the utilisation of such systems makes emergency managers increasingly responsible for actions taken in natural disasters.

4.2 The Distribution and Delegation of Morality

4.2.1 Distributed Morality

Before proceeding further it is useful to outline Floridi's theory of Distributed Morality, a concept that plays a very important role in IE, and one which can aid in contextualising ICT systems that process raw data on social media into actionable information (for the purposes of good or evil, as the case may be) as loci of morally loaded action.

Floridi (2013, pp. 262-267) argues that not all actions pass the moral threshold, that actions which may be executed with good or evil intent (potentially good or evil actions) may not always (and in fact will mostly not) have a meaningful impact on the infosphere—these actions are morally negligible (or neutral) because they are value free, "insufficiently morally loaded", or are off-set by corresponding actions. In addition to this, morally negligible actions may fail to pass the moral threshold where environments are morally resilient (that is, the environment has tolerance for or is resilient to evil actions) or morally inert (which is the opposite case, the environment is vulnerable to evil action) (Floridi, 2013, p. 266).

Of course, if one assumes that any given environment has a certain level of either moral resilience or inertia, it will have varying levels of vulnerability to good or evil actions—the case may be that, as Floridi (2013, p. 267) argues, it is only through the result of aggregated or combined individual acts that either resilience or inertia is overcome and a moral difference is made. Evilly charged actions may interact towards an evil outcome, however for Floridi (2013, p. 269), the challenge of DM is harnessing its power so that morally negligible and disparate individual actions can be channelled into one large morally good action. According to Floridi (2013, pp. 269-270), a route towards this goal involves management of moral resilience and moral inertia through policies of:

- Aggregation of possibly good actions.
- Fragmentation of evil actions so that they might be isolated and neutralised.
- Incentives and disincentives.

- and "technological mechanisms that work as 'moral enablers'".

Optimising DM is something then that requires a multi-faceted approach; it requires the management not only of environments but of agents, artefacts, and the interactions of all things.[47]

4.2.2 Delegating Morality to Artefacts

The delegation of morally loaded tasks to artefacts is not unusual, and can take numerous forms. The delegation of morally loaded tasks is naturally a popular topic in computer ethics as a whole, and has inspired much research (Alison Adam, 2005; Magnani, 2005; Turilli, 2007; Magnani and Bardone, 2008; Turilli and Floridi, 2009; Floridi, 2013).

In describing the moral nature of autonomous artefacts, John Moor (2006) gives two examples of such 'things' that perform towards ethical goals, implicit ethical agents and explicit ethical agents.

Implicit ethical agents are artefacts which implicitly support ethical behaviour—they are designed to operate within specific ethical parameters, constrain unethical action or support ethical outcomes (Moor, 2006, p. 19). Moor (2006) offers the examples of an ATM and auto-pilot controls in aircraft as implicit ethical agents, artefacts which generally, when functioning correctly, support morally good outcomes (the legitimate and accurate withdrawal of funds from one's account or the safe flight and landing of a plane).

Explicit ethical agents are by Moor's (2006, p. 20) description a much rarer class of artefact that make explicitly ethical decisions on the basis of ethical knowledge. Such a class of entity may be considered a responsible moral agent with some limited capacity to make decisions based on an approximation of intentional states. For the moment,

[47] Such an outlook is not necessarily new; DM is similar to Actor-Network-Theory (ANT), which also includes artefacts into the field of morality, and proposes that morality is distributed throughout interconnected things that comprise a network capable of moral action (Wiegel, 2010, p. 206; Simon, 2015, p. 154). In the case of ANT, the entity attains morality because it compromises a moral network, whereas IE does not deny that an artificial agent can be a source of moral action itself. However, for the time-being, it is sufficient to adopt a high LoA that examines the impact of an multi-agent system (MAS), a system compromised of multiple agents, in order to understand how an MAS can be shaped and influenced into being a "good MAS". In this chapter, the role that non-human entities play in constructing an effective, moral MAS will be examined.

technology may not be sufficiently ripe to classify anything in this way, though it is beyond the remit of this research to analyse this in any depth.

In broad agreement with Moor, Jos de Mul (2010, p. 226) argues that artefacts can be delegated with morality, examples include: "...implementation of moral values and norms in the design of artefacts, delegation of moral means to machines, and delegation of both moral means and goals to machines." An artefact delegated with values and norms as well as moral means could be classified as an implicitly ethical agent and it is not difficult to uncover examples. The Virtual Private Network (VPN), for example, is a software artefact that is imbued with moral values (privacy) in its very design and is delegated with moral means to protect privacy as it encrypts data sent over networks, hiding it from potentially prying eyes, and hides the IP address of its user as they browse the internet.

Magnani and Bardone (2005; 2008) argue that artefacts can be moral mediators, that they can externalise ethical knowledge, can mediate tasks by representing a problem, and by making the solution clearer. Magnani and Bardone (2008, pp. 104-105) offer the example of the website costofwar.com to support their case. Costofwar.com is an external resource that contextualises the money spent on the war in Iraq by the United States and by showing alternative uses to which it could be put. Of this site, Magnani and Bardone (2008, p. 105) argue that: "... [it] uncovers and unearths certain information that otherwise would have remained invisible or unavailable for making sound judgements... Now, we contend that the website can be considered a *moral mediator*, because it mediates the task changing representation we have of it, and making the solution more transparent."

While morality can be inscribed into artefacts, these artefacts do not necessarily operate in a vacuum free of human interaction. For that matter, artefacts that are inscribed with morality still do not necessarily promise moral outcomes. One can, for instance, use a VPN to conduct illegal business online without accountability—the VPN may have been inscribed with the moral value of privacy but that is not to say that it cannot be abused. The human element remains, artefacts can constrain, support or enhance action, however design and implementation by human agents are critical in determining how (morally) effective artefacts can be.

Alison Adam (2005, p. 223) provides an excellent real-life example of an artefact delegated with a moral task that failed to make a moral difference due to a failure of effective implementation. Adam (2005, p. 223) describes a situation where the failure of authorities to mobilise and share the contents of a database resulted in the employment of sex offenders at schools resulting in death—she describes this failure as such:

> The database does not work on its own—the whole moral network of database plus police and/or social workers, education and health officials, those who could have kept the data, passed it on, interrogated it and shared it, failed to work. So it is not enough to delegate aspects of morality to a database, the morality of the network must be distributed through human and non-human agents.

Morality then can be delegated to artefacts, but the network must be strong—DM can fail if entities, human and artefacts, cannot interact, or do not do so appropriately. In the above case, all entities involved failed to unite actions towards a common goal that would pass the moral threshold of good. The database, as a moral enabler, was present, and could have played a pivotal role in contributing to a morally good action. The moral significance of an artefact then may in cases be contingent on the quality of relations in the network (or MAS) of which it is apart.

In the following section the Slándáil EMIS will be described in detail, which is itself an artefact delegated with a moral task. The subsequent section will examine how the delegation of morality to an artefact such as Slándáil, under the right conditions, can greatly assist in using the power of DM to contribute towards large morally good actions, and in particular, by helping us escape from the tragedy of the Good Will.

4.3 Components of the Slándáil System

4.3.1 What is the Slándáil System?

The Slándáil EMIS is a system designed to harvest and process information from social media sources into actionable information for emergency managers during natural disasters. The system is the output of the EU FP 7 funded project Slándáil. The Slándáil project is a collaborative European project, and the full system is the product of the work of academics and professionals located in Ireland, Northern Ireland, Italy, and Germany.

The project was spearheaded by the School of Computer Science and Statistics in Trinity College Dublin, where much work was done on text and image processing. The Irish

School of Ecumenics lead research on the ethical implications of the system. The Garda Siochana collaborated as end-users in what was a participatory design approach.[48] Collaborators in the private company Stillwater Communications were involved in research and guidance on communications strategies in emergencies. Collaborators in private company Pintail provided project management support.

In Northern Ireland, collaborators in Ulster University conducted research on image analytics and legal research (internet law). The PSNI were also available to provide end-user feedback.

In Germany, collaborators in the private company CID worked to adapt their online media analytical tool, Topic Analyst, to process and analyse social media data. Collaborators in Institut für Angewandte Informatik at the University of Leipzig (INFAI) conducted research in linguistics and terminology, and legal research on copyright law.

Finally, in Italy, collaborators in private company Datapiano worked to adapt their EMIS, SIGE, to accommodate and integrate the social media analytical tools developed by other collaborators. In the University of Padua, research was conducted on linguistics, terminology and human rights law.

In what follows, the artefacts that comprise the total Slándáil EMIS will be described and it will subsequently be argued that, functioning together towards the same goal, the system demonstrates agency.

4.3.2 The Social Media Monitor

The Slándáil Social Media Monitor (SSMM) is a text and image processing application that at early prototype stage ran on a backend system based on a text-analysis tool called CiCui. The SSMM at present plugs into Twitter's public API and collects tweets in the backend based on geographical queries (that is, it collects all tweets that will be supplied through Twitter's API, which will not supply *every* tweet). All tweets collected are supposed to originate from within a particular geographical boundary. Twitter will supply the tweets based on the geo-location of the user[49] or their stated location.[50]

[48] Eliciting feedback from the emergency managers who would prospectively be using the final system and would have particular insight into the needs of emergency managers in disaster situations.

[49] That is, the geo-coordinates given by the device the Twitter user is using.

[50] Locations can be stated on Twitter, for instance, and it has a feature that allows a user to "check-in" to certain locations.

Work is also being done on inferring user location (such as by examining the locations of friends/followers of the Twitter user).

The backend of the system performs analysis of the text content of the tweets using natural language processing (NLP) methods. The system will process and filter messages based on relevant terminology relating to natural disasters that was added to a terminology database. Collected tweets are saved on a server to improve future system performance. The backend system can also propose candidate terms based on analysis and send them back for human review, where they can then be added to the database to improve it.

The system filters out irrelevant messages, or noise, based on this analysis and displays relevant messages (signals) to emergency managers. These messages will contain terminology from the terminology database. Messages can be geo-located on an interactive map, and can be tagged with the output of sentiment analysis, that is, the contents of the messages can be labelled as either being positive or negative in sentiment.

In terms of image analysis, the system uses Alchemy AI at time of writing, however it is expected to use a more specialised tool called C2 (developed by Ulster University primarily) when it is ready for deployment. The image analysis tools operate similarly to the text analysis; it analyses images (and any associated text) in the datastream, classifies them and tags them, displaying the relevant images (which can be geo-located on an interactive map) to the emergency manager. The image analysis tool would have been trained to recognise relevant image features from manual training—it would have essentially been 'fed' relevant images in order to recognise relevant images. Training images were collected from major news outlets and social media.

The SSMM also displays line graphs that visualise trends; these graphs are based on different categories based on the dictionaries used, including sentiment, and intrusion (which will be described in more detail in the following).

Although the SSMM is formally compatible with Twitter, efforts are being made to adapt it to the Facebook API.

The processes described are illustrated in Figure 6.

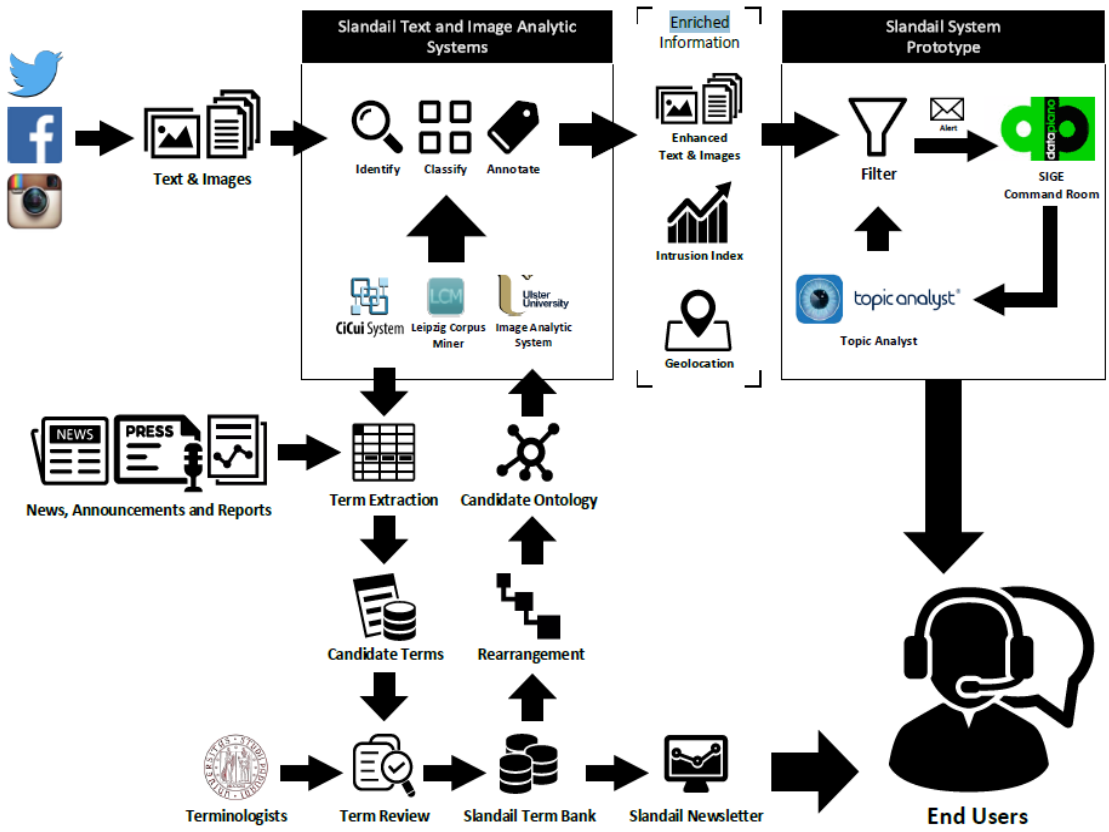

Figure 6: Dataflow of the Slándáil Emergency Management Information System (Source: Slándáil-TCD, 2016)

A follow up interview was held with the relevant expert and the SSMM was updated. The SSMM was rebuilt from the ground up and its most recent iteration does not utilise CiCui, although its backend does use some of CiCui's code. This was done for efficiency as not all aspects of the CiCui system were utilised in the initial prototype, it was described as being "...trimmed down to essentials," but "...[t]he functionality is pretty much the same...". This was done for the purposes of integration and interoperability, better accommodating Ulster's image analysis algorithm and providing for the possibility of integration with non-SIGE EMIS. This iteration moved away from the presentation of an exclusive interface, with the intention that the presentation of data can be adapted onto the EMIS with which it works in conjunction.

4.3.3 The Intrusion Index

The intrusion index (II) was devised as a computational method to discern how much sensitive data is being processed by the SSMM. It is built into the SSMM and functions much the same as the detection of relevant disaster related information, only in this case its dictionary is based on named entities such as people, places and organisations

(essentially pronouns)—this method is named entity recognition (NER). The II records the frequency of occurrence of named entities over time and visualises this data on a graph (see Figure 7). One can anticipate a peak in the line graph where an incident is occurring.

The II will also flag potentially sensitive information on tweets that are pinned to the interactive map, and colour these pins based on content.[51]

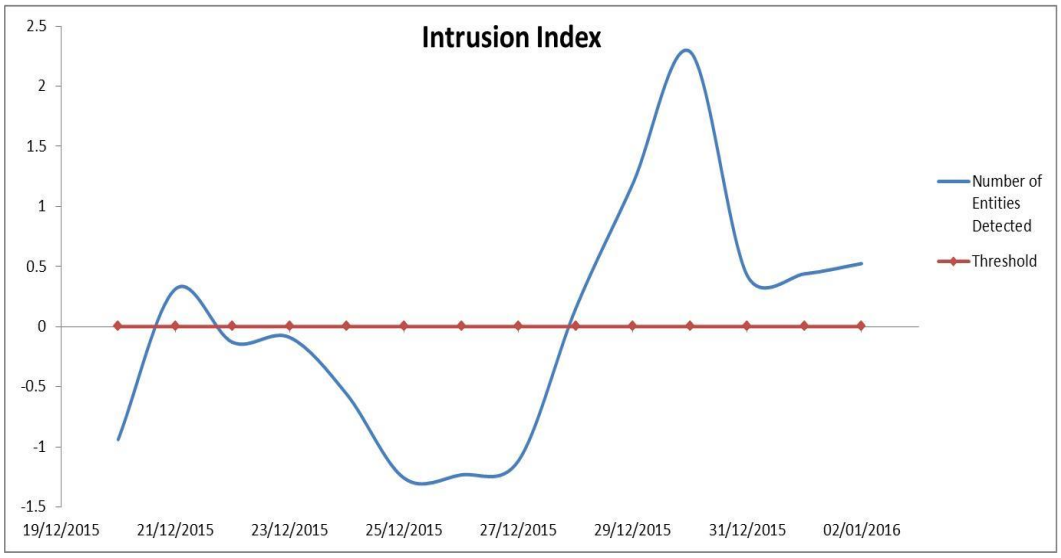

Figure 7: Line Graph Illustrating Frequency of Occurrence of Named Entities over Time (Source: Slándáil-TCD, 2016)

[51] See Figure 8 for an example, where tweets containing sensitive information have blue pins, and tweets without sensitive information have yellow pins.

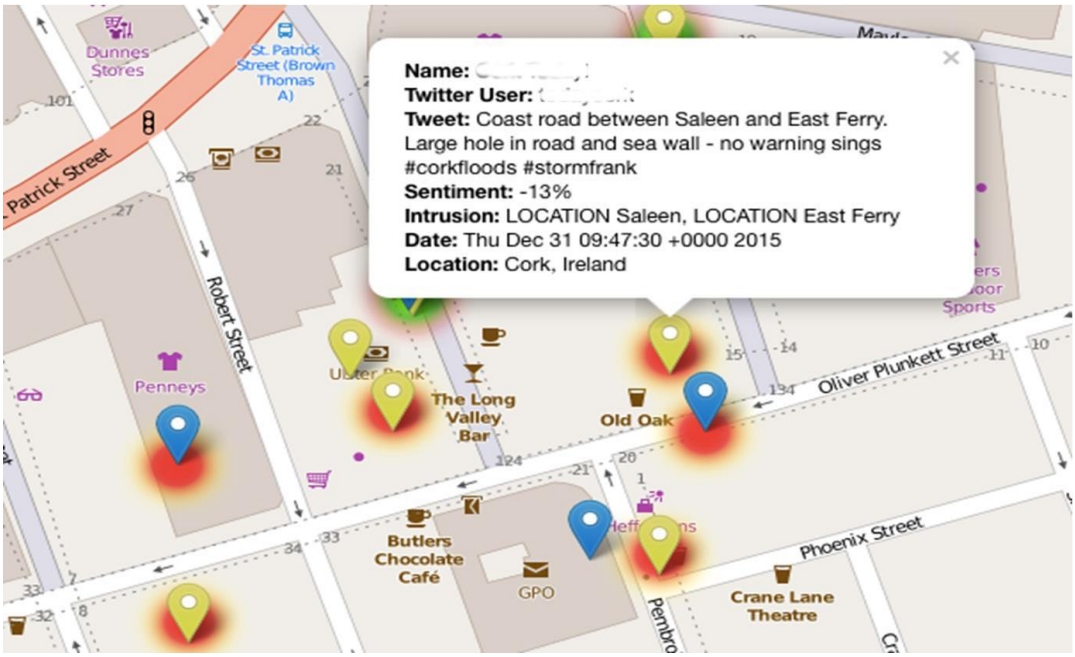

Figure 8: Visualisation of Tweets on Map Including Intrusion Information (Source: Slándáil-TCD, 2016)

4.3.4 Statistical Modelling for Decision Support

Another computational artefacft proposed to be added to the final iteration of the system is a Bonferroni mean aggregation model.[52]

The model functions by measuring the relevance of social media content in an emergency event through an index and rolling this data into a model that includes multiple other variables (including geo-spatial information). The variables in the model are intended to include a wide range of location related information, for instance area use (whether there are a lot of buildings), population, and rainfall data.

The variables' interactions are weighted and the output of the model is proposed to be a three tier scale that will indicate the severity of risk in any given area that the emergency manager queries on an interactive map. This score is proposed to be displayed on the map, and can be recalculated when a different area is queried.

4.3.5 Topic Analyst

Topic Analyst is a system that analyses digital media and which has recently been adapted to analyse social media messages. Topic Analyst is a more visual platform than the SSMM, providing numerous analytical visual features including word clouds and pie

[52] This was at the theoretical stages of development and not implemented at time of interview.

charts that detail "hot topics" or frequently mentioned terms in collected documents and the proportion of occurrence of terms.

Topic Analyst enables the monitoring of documents (including tweets) published online by topic, by creating a topic to monitor and entering relevant seed words. The system filters out documents unrelated to the seed words and presents all documents that it deems relevant. Topic Analyst uses web-crawlers for the discovery of documents and NLP to analyse these documents.

In these documents, metadata and raw text are viewable from within the system's dashboard. The user can follow a link to the original document as it appears online, including at present, tweets. Metadata and accompanying text (that is, document content) can also be exported and saved in a .csv format. Figure 11 shows retrieved social media messages relating to a topic selected by the researcher, in this case, simply "disaster". At time of writing, the names of tweet authors are omitted from results displayed within the dashboard, though @ mentions to other users are visible. The system also has geo-spatial capabilities, and can display the number of documents published in each location in clusters, where geo-tagged data is available. Topic Analyst is capable of processing documents from not just Twitter, but other social media sources including Facebook, Youtube and potentially Google+.

Topic Analyst allows in depth exploration of topics, and will display words or terms that co-occur with a search (See Figure 10), so for instance, if the queried topic is flood and a co-occurrence is bridge, an emergency manager may be able to explore documents related to the co-occurrence and discover that a bridge has become impassable. It also allows for users to create alerts based on supervised topic, therefore an emergency manager for instance can be notified if there are any updates in a supervised topic monitoring documents pertaining to a flood, which can be pushed through to the Slándáil system.

Topic Analyst can import data obtained from the SMM and use its analysis on tweets collected by its backend. Topic Analyst has broad topic monitoring capabilities but to apply analysis to emergency events on more specific and localised scales it depends on what it receives from the SMM backend, and thereby the dictionaries already in place that enable filtering.

Whilst Topic Analyst has very broad, global even, document analysis capabilities (by default it detects popular topics and trends), it can operate on a more localised scale and provide analysis of documents at a regional or city level.

The system uses machine learning in the backend to improve performance and its NLP capabilities, including entity recognition.

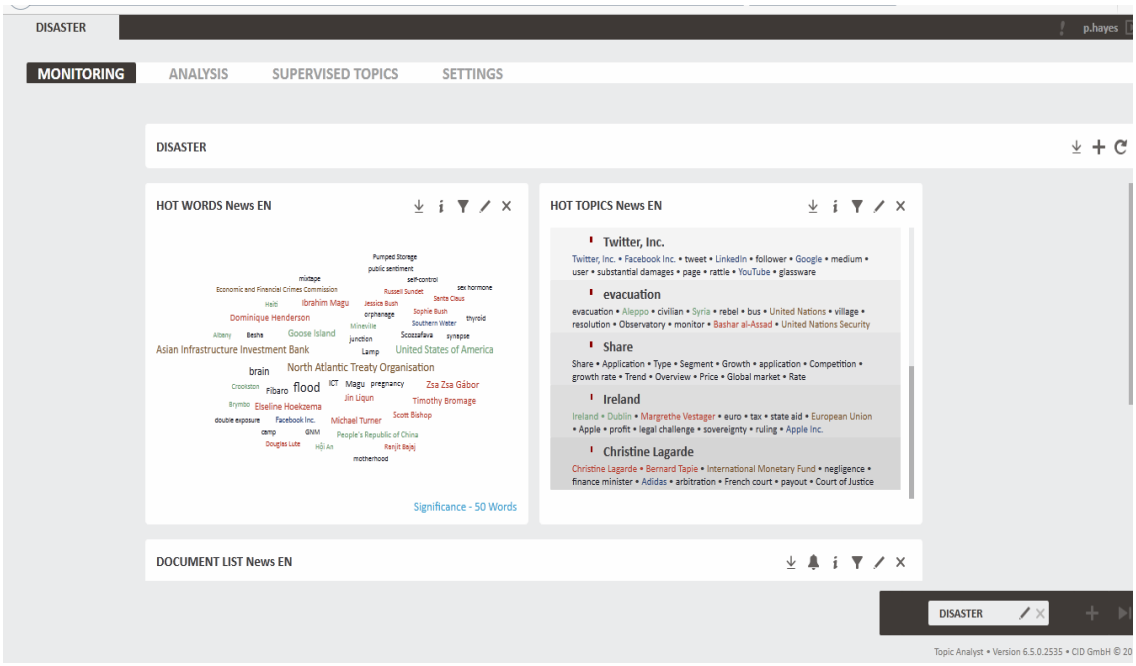

Figure 9: The Topic Analyst Dashboard Showing Hot Words and Topics (Source: CID, 2016)

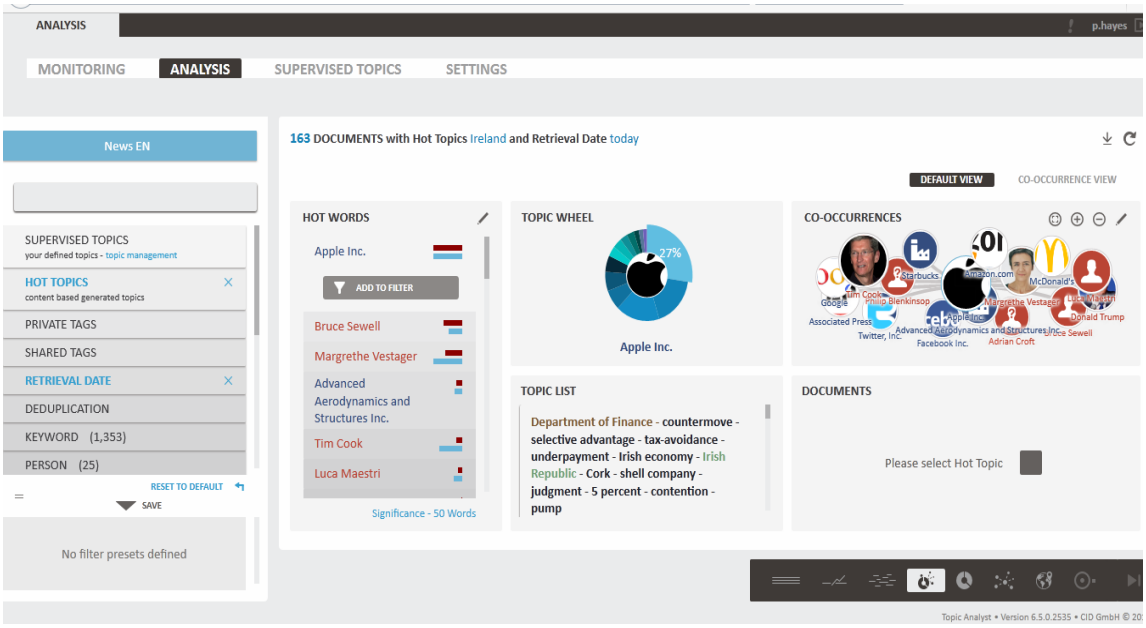

Figure 10: Topic Analyst Analytical Features (Source: CID, 2016)

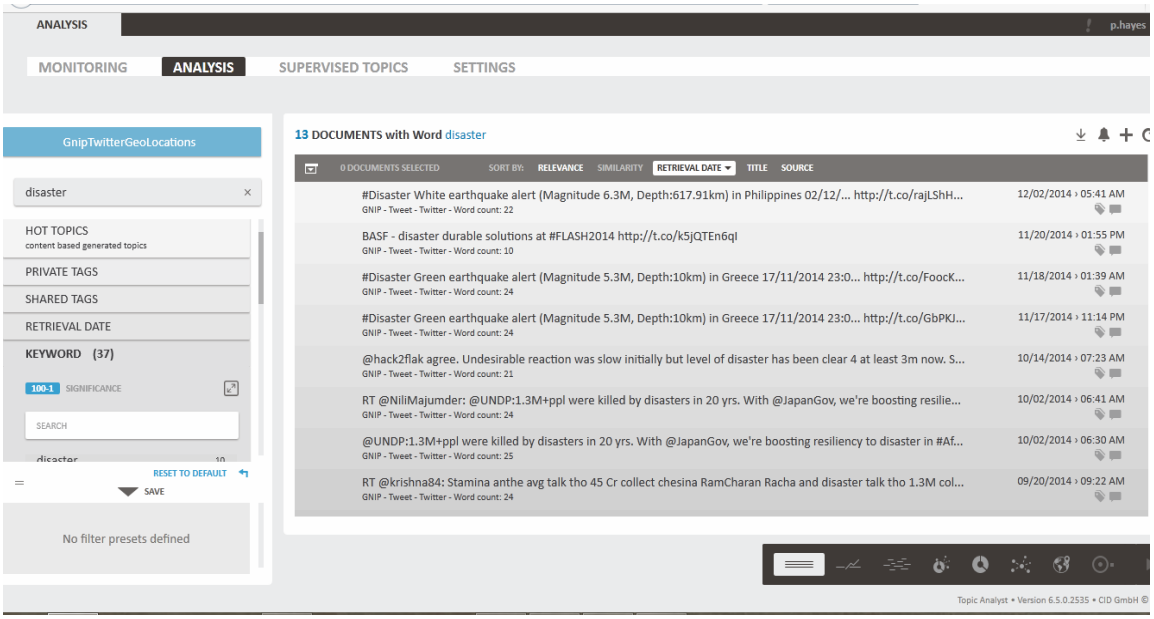

Figure 11: List of Tweets Relating to a Topic (Source: CID, 2016)

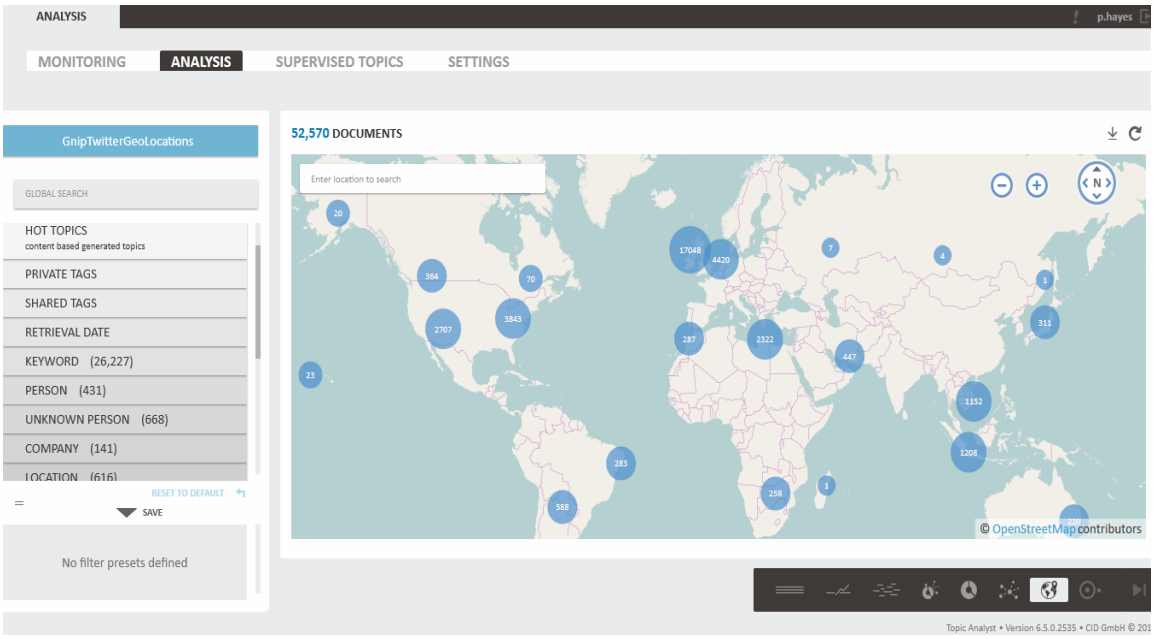

Figure 12: Geo-location of Tweets (Source: CID, 2016)

4.3.6 SIGE

The SIGE system is the EMIS in which, metaphorically, the other system components are wrapped, and it is the first interface the emergency manager will see (see Figure 13 to see SIGE's dashboard). SIGE is a traditional emergency management tool that is being

updated to integrate or at least provide a portal to the social media functionality of the SMM and Topic Analyst.

SIGE currently provides a portal to the other central components, which are at present separate but linked platforms, but with further integration, can function more coherently as one more unified, complete package.

SIGE was described in interview by those working on it as being akin to an "instruction booklet" for emergency managers. It is a highly configurable tool that can allow users to create events and list processes and tasks to follow which are tailored to specific events, such as a flood or earthquake (see Figure 14). In essence, an event can be configured to list the appropriate actions to take at any stage in an emergency, it functions as a kind of reminder or checklist of possible activities.

The most important functionality of SIGE is its Geographic Information Systems (GIS) functionality, which is an interactive map capable of supporting and displaying a range of data as overlays. Any data that is available pertaining to any boundary unit can be displayed on the map, although at present, only information that DataPiano has direct access to can be converted into shapefiles for use with the map.[53] DataPiano are working on a feature to facilitate the importing of data by end-users.

In Figure 15 see an example of the interactive map, which is focussed on a portion of a town in Italy. The map can display facilities and buildings that are in the area, including buildings such as hospitals that would be essential resources in the event of an emergency. Also observe on the left column that there is a range of other data that can be overlaid including Hydrography. In Figure 16 the hydrological risk map layer has been selected, displaying areas of the map that are vulnerable to flooding.

It is intended that social media messages filtered by the SMM will be available as a map layer within the SIGE platform itself, however the live importing and visualisation of data as a process integrated within SIGE has yet to be completed at time of writing.

The SIGE system does not by default store any social media information, however a user can define information that they deem important to be saved.

[53] Such as on the INSPIRE database where EU states are mandated to keep record of publicly available data.

It is the preference of DataPiano that the entire architecture of the Slándáil system be stored on the premises of emergency managers in a secure environment—which is to say that the software be installed locally rather than accessible remotely through the internet.

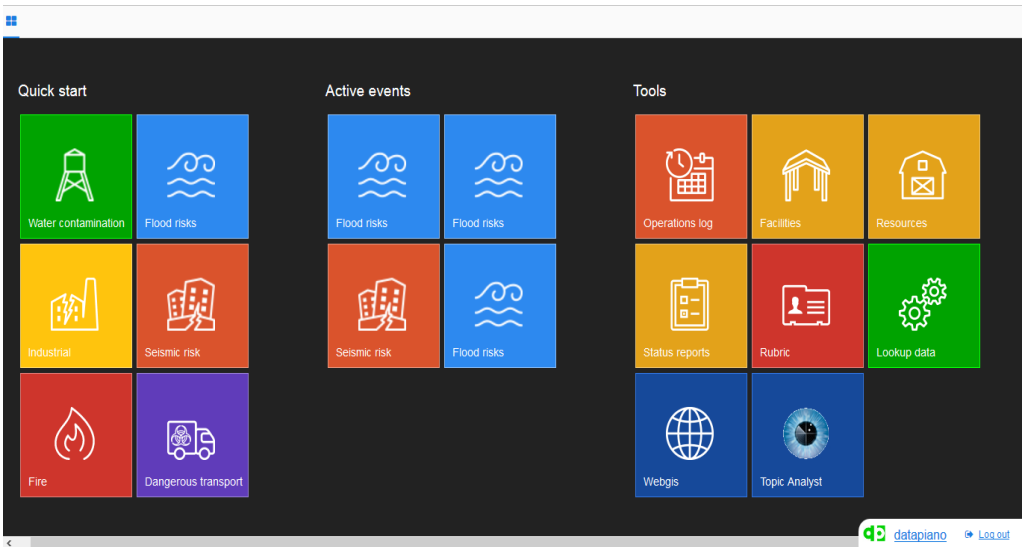

Figure 13: SIGE Dashboard (Source: Datapiano, 2016)

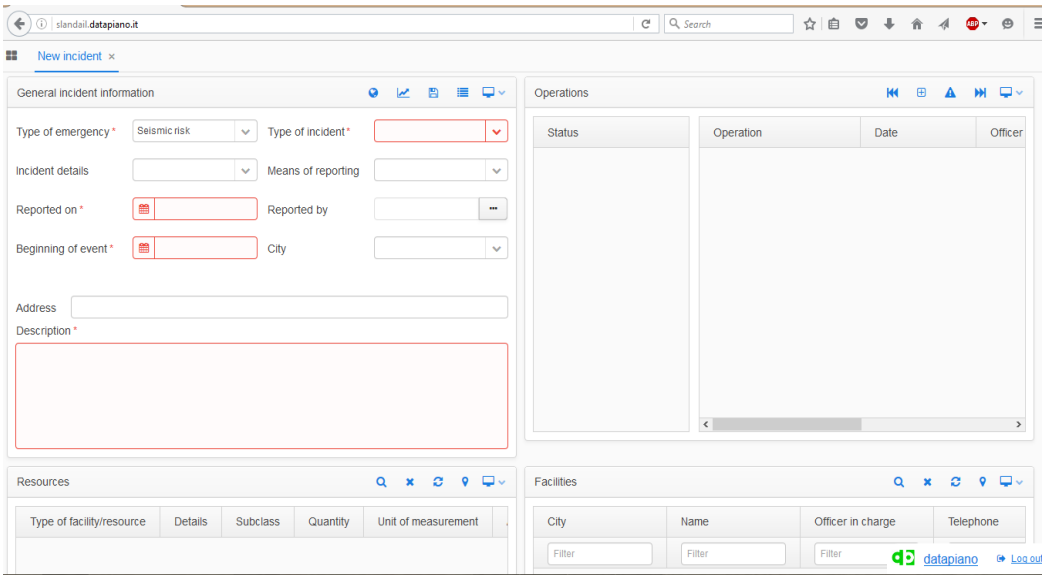

Figure 14: SIGE Incident Creation Feature (Source: Datapiano, 2016)

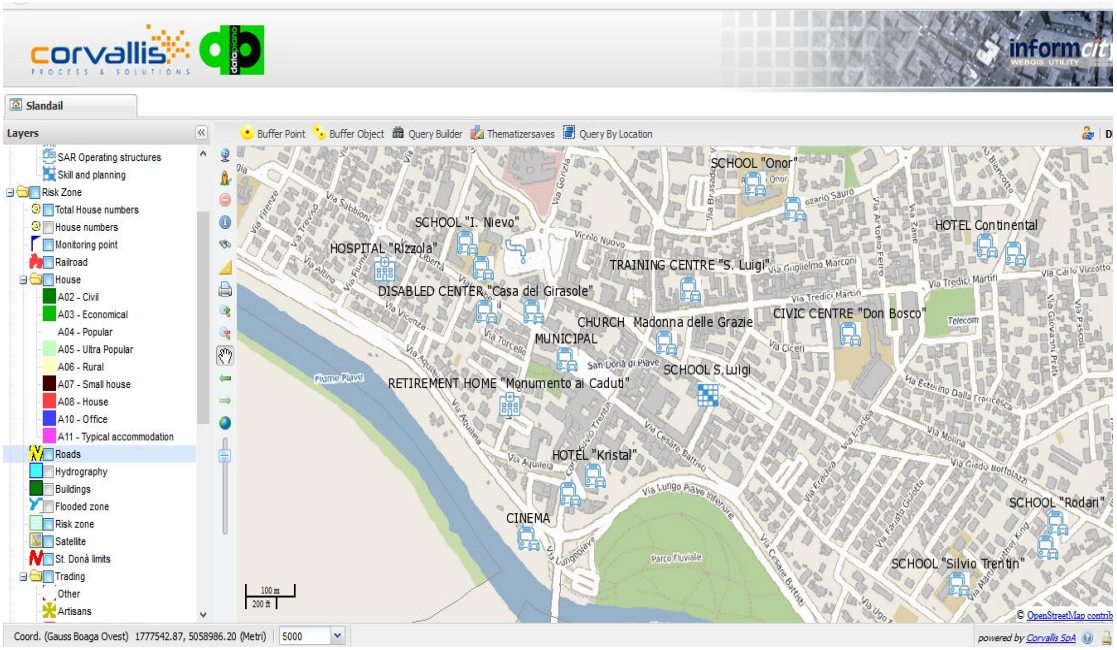

Figure 15: SIGE GIS map Featuring Layer of Local Resources (Source: Datapiano, 2016)

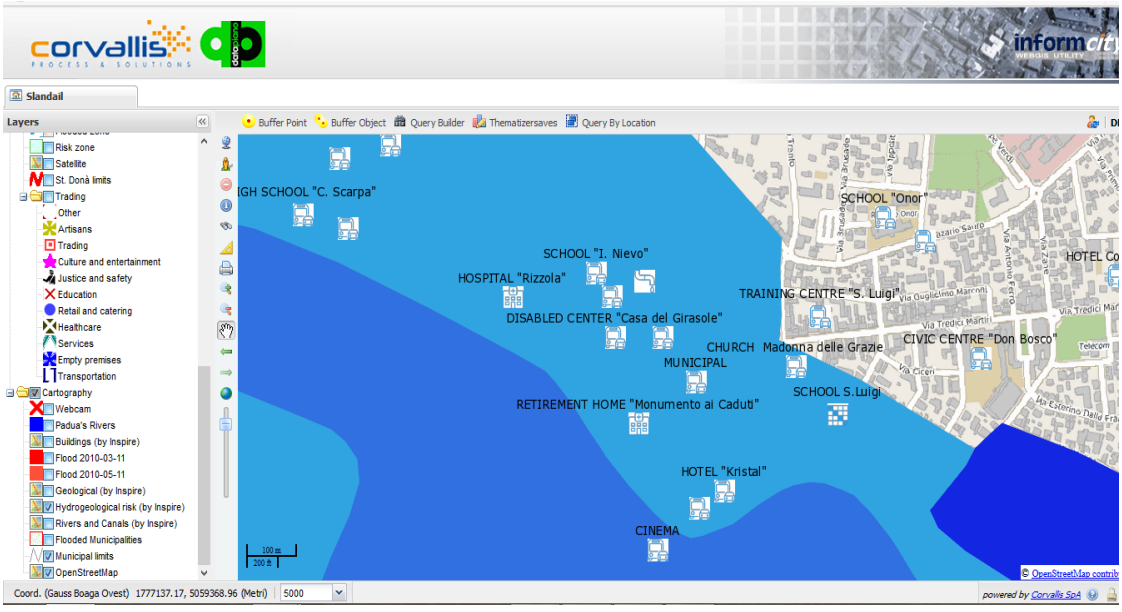

Figure 16: SIGE GIS Map Featuring Layer of Local Resources and Hydrological Risk (Source: Datapiano, 2016)

4.3.7 System Configurations

At time of writing, the system was not finalised or commercialised and therefore its future and exact configurations remain uncertain. This is to some extent advantageous

to the researcher, who is free to posit different prospective configurations of the system.

The system is the sum of its parts as outlined in this chapter, therefore it should be noted that nearly any individual component can be used here exclusively, where possible, or in conjunction with other compatible software. For instance, SIGE and Topic Analyst may be utilised together but other artefacts developed throughout the Slándáil project could be omitted. Likewise the SSMM could be adapted to non-SIGE EMIS to approximately the same effect it would have within SIGE. The use of algorithms such as Ulster University's image analysis algorithm could potentially have rather open-ended use. This is to say that the pieces of the whole could be used separately, or within the same context of the overall system described here but with certain components omitted or swapped out for similarly functioning components (other software artefacts that provide the same functionality but were developed/licensed by different entities).

The research will proceed generally assuming that the artefacts developed or adapted under Slándáil will function together. This would mean that all the overviewed artefacts will function together and will be used by an emergency manager; the EMIS would support social media analysis through the SSMM, which can analyse text and images (using either Alchemy or C2), and data would be passed along to Topic Analyst for further analysis.

These components can technically be stored in and operated from the premises of an emergency manager, or operated remotely from the premises of their creators. Whether CiCui would be utilised remains to be seen with the addition of the updated SSMM, but it is not an impossibility. Regardless, both artefact iterations have similar functionality.

4.3.8 Slándáil, Agency and Delegated Morality

The Slándáil system is an example of the delegation of a morality to an artefact. The system is charged with the morally loaded task of processing and filtering data on social media in natural disaster situations—a task that can be at best challenging and require significant manpower in high population areas where tweets will be generated in high volumes. Emergency managers have a need not just for data in such situations, as an interview participant from the emergency services noted, they are not experts in any given domain, and need this data to be processed into actionable "intelligence", which is

to say they need data to have a level of interpretation applied to it. In filtering relevant messages and geo-locating the messages on an interactive map (in the ideal implementation scenario), the system succeeds in doing this by displaying actionable intelligence—the emergency responder will potentially be presented with a message of distress and an exact location with a reduced burden of interpretation. Combined with the existing capability of SIGE's GIS functionality, the emergency manager can add map layers of processed information displaying anything from hydrological information to the population of a given area to further assess risks to individuals in emergency situations. With the realisation of the Bonferroni aggregate model, further inferences or interpretation can be applied to available data and decision making can be further supported by the delegation of a task to this artefact—this method can weigh up relevant data and advise emergency managers which areas are at particular risk. Furthermore, the addition of further social media data analysis by either the SMM or Topic Analyst for instance can indicate particular discussion trends within a disaster, potentially indicating urgent situations (for instance, where "flooding" and a particular "bridge" are trending).

This example of delegation to these particular kinds of tasks can be considered "cognitive delegation" (Magnani, 2005, p. 5). The system represents a problem, and makes the solution to that problem easier to discern (Magnani and Bardone, 2008).

Here, for the sake of simplicity, the system has been referred to as an artefact,[54] though it is at this stage worth assessing if the system can be considered an agent itself.

At perhaps multiple levels of abstraction the system fails to meet the demands of agency by not meeting the criteria of autonomy, interactivity and adaptability. The adaptability criterion is a difficult one when applied to IT artefacts, as often any form of adaptation will be supervised or implemented by a human agent. For the software developers operating behind the scenes of this system, it will not be an agent. The system will propose candidate terms to be placed in a database to improve it, but these terms must be approved by humans—it fails to be adaptable.

For the emergency managers though, over a period of time, the total system could be classified as an agent. It is autonomous inasmuch as it automatically carries out the tasks of data collection, processing and analysis; it is interactive as the user can interact with it

[54] Though it is probably more accurate to call it, in the words of Turilli and Floridi (2009), an autonomous computational artefact.

in numerous ways (for example, setting search parameters), and it is, over time, adaptable, as it is continuously improving in its filtering capabilities potentially without input from the emergency manager. Regardless of LoA, the system, and its components are artefacts that have the potential to make morally charged contributions to action in the infosphere.

The system cannot be considered a moral agent; however, its output is dependent on human agency for any morally good action to be executed. At best then, the system is a moral enabler that enhances human agency. This is not to undermine the potential that this system has to improve (or undermine) a multi-agent system, and in the next section its capacity to contribute to morally good actions with respect to its place in a multi-agent system will be assessed.

4.4 Life and the Tragedy of The Good Will

4.4.1 Harnessing the Power of Distributed Morality to Escape the Tragedy of the Good Will

Having outlined the key points of functionality of the Slándáil system, an analysis of its potential benefits will be undertaken before embarking on the task in the following chapters of analysing its implications for the values of privacy, justice, trust, responsibility and accountability. Here, it will be argued that the system, as an artefact that has been delegated morality, has the capacity to harness the power of DM and has to potential to avert what Floridi (2013) refers to as the *tragedy of the Good Will*.

Before proceeding further, it is instructive to outline what precisely the Good Will is, and then what it is that may be called the tragedy of the Good Will.

According to Floridi, the Good Will is a morally responsible agent that is ethically caring, attentive to the world around them and importantly, desires to do good in this world (Floridi, 2013, p. 197). Floridi (2013, p. 197) argues that the Good Will is:

> ...endowed with some albeit limited resources, who bases her decisions and actions on the proper management of her factual information about the moral situation in which she is involved..., who is reasonably capable of implementing whatever she thinks ought to be done morally, whose responsibilities increase with the amount of information she enjoys (and who knows that this is the case),

and who is motivated by a genuine desire to know and by a sincere *eudokia*,[55] while not suffering from *akrasia.*[56]

The tragedy of the Good Will arises when there is a fundamental imbalance between the Good Will's information regarding a particular moral situation and their power to influence this moral situation, to perhaps implement an action that can prevent an evil from occurring (Floridi, 2013, pp. 197–198).

Suppose that a river were to break its banks in a small town during an intense precipitation event, inundating a large proportion of the surface area of the town and trapping people in their homes. The waters are high and fast moving; many are trapped on roofs and first floors. In this situation we may have two agents representing Good Wills, the responsible emergency response agency and Alice, a woman affected by disaster who is trapped in her home (also a patient, as a recipient of the forces of natural evil and a potential rescuee). Both Alice and the emergency response agency seek to do good and eliminate the deleterious effects of this natural evil. However there is a problem; Alice has firsthand knowledge of the situation but being trapped in her flooded home has no power to help others—from her bedroom window she can see neighbours standing at their balconies signalling for help. She does not have the resources to aid them, she requires help herself. Her neighbour's house looks unstable, making the situation even more urgent and precarious. Alice calls the emergency operators but the lines are busy, so she does what little she can and takes a picture on her android phone and posts it on Twitter.

The emergency management agency has the opposite problem, it has the power to mitigate the effects of this natural evil; it can deploy personnel to the location to rescue both Alice and her neighbours, however in the midst of the confusion and under a deluge of noisy social media and phone reports from affected citizens, it does not have all relevant information concerning the moral situation. Time passes and Alice's neighbour's house collapses into the water. Everyone perishes. The tragedy of the Good Will has manifested—Alice, with the will to do good but not the power to change anything, witnesses her neighbours succumb to natural evil, whilst the emergency management agency with the will and the power to do good has incomplete information regarding the moral situation and is unable to use its power in time to save Alice's neighbours.

[55] According to Floridi (2013, p.196), *eudokia* is a "...willingness/desire to do the right thing....".
[56] Where through one's weakness of will, they act against their better judgement (Stroud, 2014).

In the example described, Alice committed to a morally loaded action, however due to the moral inertia of the environment, this action (posting an image on Twitter) was insufficiently morally loaded to rise above the moral threshold and make a meaningful difference. The tragedy arises as Alice has knowledge of a moral situation without the power to affect it, and the emergency response agency has the power but insufficient knowledge to affect this particular situation.

This example not just highlights an example of the tragedy of the Good Will, but demonstrates that in a natural disaster situation the asymmetry of knowledge and power between disaster survivors and emergency manager can result in life-threatening inefficiencies.

Floridi (2013, p. 204) argues that there are four ways whereby the tragedy can be escaped:

1. the information/power gap may decrease, as information has already reached its peak, whereas power is catching up;
2. from quantity to quality of information: better informed Good Wills can act and exercise their augmented power better;
3. from the powerless observation of the single Good Will to the empowered interactions of multi-agent systems of Good Wills: global problems and distributed morality require global agents;
4. the ontological side of information: the need for an augmented ethics.

It is argued here that the Slándáil EMIS can address the first two of these four points (and partially the third), whilst the central premise of this research is that an augmented ethics, IE, can better help understand the challenges of modern ICTs and effectively by doing this can outline solutions—this chapter itself represents the first implementation of the theory here in order to expose the challenges relating to ICTs in emergencies, and solutions.

The Slándáil EMIS serves as an example of a system that has been delegated a moral task of processing social media data during emergency events and one which stands to help harness the power of DM towards escaping the tragedy of the Good Will. Slándáil can reduce the information/power gap by aggregating the individual morally negligible actions of agents such as Alice, by filtering their messages (by recognising that they are signals and abstracting the noise), pinning them to an interactive map, and making them more available to emergency managers to gain enhanced situational awareness (or information pertaining to the moral situation). Pre-existing information can be overlaid on these maps, for instance hydrological information or flood risk information, which

could potentially confirm the seriousness and validity of the social media messages being presented—or means such as Bonferroni weighting could numerically display the risk by weighing social media messages and other data in aggregate form. Trend analysis of social media data might also succinctly indicate areas under severe threat.

Slándáil, as a kind of moral (or cognitive) mediator, represents a problem and thereby makes the solution to that problem more transparent. It is empowered by individual participation, and it empowers individual participation. Through this represented information (previously lost data in a noisy datastream), emergency managers sitting in operational centres may be sufficiently informed to deploy resources where they are most needed that can either confirm or act in the situation.

Slándáil reverses the situation of Alice so that she is not merely a passive patient, but is actively participating in an MAS, one which is actively self-repairing and eliminating sources of entropy arising from natural evil due to the managed and enhanced interactions between its constituent agents—the Slándáil system acts as a mediator that bridges the gap between agent and patient, bridges their strengths and weaknesses and supports collective action. By connecting Alice's actions with those of others, it reduces moral inertia in the environment, and envelopes Alice's formerly morally negligible actions into a system with positive, morally meaningful output—it empowers the passive patient to be an effective morally responsible agent (or Good Will).

All this is not to say that a system such as Slándáil will aid in escaping the tragedy of the Good Will every time, nor that it would not be subject to technical or other faults (recall Adams' example and how each agent in the system must be acting in the right way, and under the correct parameters for the good to occur). However, it does demonstrate the capacity of an artificial agent to manage interactions in an MAS to aggregate actions, represent problems and inform and support decision making by those in a position to utilise power to implement meaningful actions that can eliminate sources of entropy in the infosphere. In this situation, the credit does not go to any one agent, but all those who were a part of the network leading to a morally good action. Systems such as Slándáil of course have the raw potential to help avert the tragedy of the Good Will, but are not guaranteed to. Their efficacy is based on an ideal scenario of sufficient available rescue resources, sufficiently available social media data, and environmental hazards that are not insurmountable. Nonetheless the technology presents very attractive possibilities in contributing to moral action. It also has the potential to likewise

contribute to evil action, and such possible contributions will be assessed in the chapters that follow.

4.5 The Human Rights Perspective: Protection of Life

4.5.1 The Right to Life in Theory and Law

If the state's fiduciary duty is to provide a regime of secure and equal freedom under the rule of law, under conditions of non-instrumentalisation and non-domination, then surely natural disasters are anathema to this duty, exposing subjects to dearth and dangers that threaten the realisation of a broad range of human rights.

Natural disasters can endanger the right to life, have resulted in catastrophic loss of life, and will likely continue to do so moving forward. The right to life is among the most fundamental that a state can protect, for without safety of the person and the assurance that their survival has some measure of guarantee, there can be no regime of *secure and equal* freedom. Protecting the right to life entails positive obligations, that is, the state should actively deter any human threats and should take all reasonably implementable actions to safeguard it (including from environmental threats). In the case of natural disaster, whilst the case may be that man can be dominated by his fellow man in the aftermath of an event that jeopardises the state's instruments from establishing control and curtailing their ability to enforce a regime of secure and equal freedom, it should also be considered that the subject has become dominated by the elements, that their agency can be greatly restricted by natural hazards and their dangerous environment. In such situations human dignity is under great threat.

The right to life as a basic concept is uncontroversial,[57] and has been established in international law. The *Universal Declaration of Human Rights* (1948), Article 3 affirms that "[e]veryone has the right to life, liberty and security of person," whilst the *United Nations International Covenant on Civil and Political Rights* (1966), Article 6, paragraph 1, states, "[e]very human being has the inherent right to life. This right shall be protected by law. No one shall be arbitrarily deprived of his life." In the *European Convention on Human Rights* (1950), the right to life is similarly protected in Article 2.

[57] At a deeper reading however, tensions may arise where one considers the application of capital punishment, euthanasia, or the question of right to life of the unborn—matters well outside of the remit of this research.

The case law of the European Court of Human Rights (ECtHR) offers guidance and elaboration on how protection of the right to life ought to be implemented by states. Positive obligations of the state to protect the life of it subjects were affirmed in *Osman v. the United Kingdom* [1998] and *L.C.B. v. the United Kingdom* [1998]. The former case concerned the murder of a young boy by an obsessed teacher and the police's failure to prevent the tragedy. The latter case concerned the exposure of the applicant's father to nuclear radiation in the Christmas Island nuclear tests, her subsequent leukaemia diagnosis and the failure of the UK Government to provide adequate information regarding the effects of radiation and monitor her health. No violations of Article 2 were found in either case as the states were found to have done all that they reasonably could be expected to do respective to the unique situations.[58] Nonetheless, the Court did comment on the state's positive obligations, affirming that states, as a matter of positive obligation, are not simply required to refrain from taking life but are required to take active measures to protect it.[59] This positive obligation extends to deterring state officials from negligent or corrupt acts that may result in the deaths of others, that is, they too must be deterred from failing negligently in their duties to prevent avoidable loss of life—the Court affirmed this in *Oneryildz v. Turkey* [2004].[60]

---

[58] On reasonable, for instance, in *Osman* [1998] the Court noted that "…bearing in mind the difficulties involved in policing modern societies, the unpredictability of human conduct and the operational choices which must be made in terms of priorities and resources, such an obligation must be interpreted in a way which does not impose an impossible or disproportionate burden on the authorities. Accordingly, not every claimed risk to life can entail for the authorities a Convention requirement to take operational measures to prevent that risk from materialising."

[59] In *Osman* [1998], for example, the Court stated that:

> The Court notes that the first sentence of Article 2 § 1 enjoins the State not only to refrain from the intentional and unlawful taking of life, but also to take appropriate steps to safeguard the lives of those within its jurisdiction.... It is common ground that the State's obligation in this respect extends beyond its primary duty to secure the right to life by putting in place effective criminal-law provisions to deter the commission of offences against the person backed up by law-enforcement machinery for the prevention, suppression and sanctioning of breaches of such provisions.

And in *L.C.B* [1998]:

> In this connection, the Court considers that the first sentence of Article 2 § 1 enjoins the State not only to refrain from the intentional and unlawful taking of life, but also to take appropriate steps to safeguard the lives of those within its jurisdiction...

[60] In this case, whereupon a methane explosion in a rubbish dump caused a landslide resulting in 39 deaths, the Court stated (Korff, 2006, p. 62 citing *Oneryildz v. Turkey,* [2004]):

> The positive obligation to take all appropriate steps to safeguard life for the purposes of Article 2 [...] entails above all a primary duty on the State to put in place a legislative and

In this case, the state was found to have violated Article 2 as it should have known that there was a threat to life, and failed to adopt adequate measures to protect it. On the procedural aspect, the domestic criminal proceedings were limited, merely finding responsible parties guilty of negligence without any substantial punishment for the guilty besides suspended fines (Korff, 2006, pp. 64–65, citing *Oneryildz v. Turkey,* [2004]). The procedural aspect of the right to life therefore requires that state officials who have responsibility over matters that may threaten the lives of subjects are held accountable should they fail in discharging their duty to protect life when it is explicitly in danger, and where reasonable means to prevent loss of life can be implemented. The procedural aspect then functions as a deterrent for state officials or authorities from ignoring their duty. This duty, the protection of life of all those subject to the state's

---

> administrative framework designed to provide effective deterrence against threats to the right to life [...].

The Court in this case also affirmed that there must be procedural requirements in place (Korff, 2006, pp. 63–64, citing *Oneryildz v. Turkey,* [2004]):

> ⋯ the judicial system required by Article 2 must make provision for an independent and impartial official investigation procedure that satisfies certain minimum standards as to effectiveness and is capable of ensuring that criminal penalties are applied where lives are lost as a result of a dangerous activity if and to the extent that this is justified by the findings of the investigation [⋯]. In such cases, the competent authorities must act with exemplary diligence and promptness and must of their own motion initiate investigations capable of, firstly, ascertaining the circumstances in which the incident took place and any shortcomings in the operation of the regulatory system and, secondly, identifying the State officials or authorities involved in whatever capacity in the chain of events in issue.
>
> That said, the requirements of Article 2 go beyond the stage of the official investigation, where this has led to the institution of proceedings in the national courts; the proceedings as a whole, including the trial stage, must satisfy the requirements of the positive obligation to protect lives through the law.
> It should in no way be inferred from the foregoing that Article 2 may entail the right for an applicant to have third parties prosecuted or sentenced for a criminal offence [⋯] or an absolute obligation for all prosecutions to result in conviction, or indeed in a particular sentence [⋯].
>
> On the other hand, the national courts should not under any circumstances be prepared to allow life-endangering offences to go unpunished. This is essential for maintaining public confidence and ensuring adherence to the rule of law and for preventing any appearance of tolerance of or collusion in unlawful acts [⋯]. The Courts task therefore consists in reviewing whether and to what extent the courts, in reaching their conclusion, may be deemed to have submitted the case to the careful scrutiny required by Article 2 of the Convention, so that the deterrent effect of the judicial system in place and the significance of the role it is required to play in preventing violations of the right to life are not undermined.

power, is a fiduciary one, and the state which rules with *integrity* must prosecute officials who fail to uphold the duties accompanying their office.

The principles emerging from the case law concerning Article 2 have been applied explicitly to the case of natural disaster in *Budayeva and Others v. Russia* [2008]. This case concerned the occurrence of destructive mudslides in Tyrnauz, Russia, 2000, that resulted in substantial loss of life and destruction of property. Again the Court affirmed the state's positive obligation to protect the right to life through preventive and regulatory measures, including measures that would enable the transmission of information to the public (such as warning of the natural hazard), as well as the procedural requirement of judicial enquiry following any deaths.[61]

The Court is lenient as regards which measures are taken to protect the right to life, in-keeping with its principle of margin of appreciation, and in this case reiterated this, also noting that a wide margin of appreciation applies in meteorological events that are beyond human control.

In the case of *Budayeva* [2008] the Court found that the state failed to meet its positive obligations to protect its subjects' lives. The threat to human life was deemed foreseeable as the threat posed by a mudslide was known; the scope of work necessary to protect life in terms of defence infrastructure maintenance and upgrades was known, however the state failed to act. The state also failed to act in informing subjects of the risks they faced, "an essential practical measure" and failed to plan for emergency

[61] For elaboration, the Court noted in *Budayeva and Others v. Russia* [2008]:

> The obligation on the part of the State to safeguard the lives of those within its jurisdiction has been interpreted so as to include both substantive and procedural aspects, notably a positive obligation to take regulatory measures and to adequately inform the public about any lifethreatening emergency, and to ensure that any occasion of the deaths caused thereby would be followed by a judicial enquiry.

And:

> They must govern the licensing, setting up, operation, security and supervision of the activity and must make it compulsory for all those concerned to take practical measures to ensure the effective protection of citizens whose lives might be endangered by the inherent risks. Among these preventive measures, particular emphasis should be placed on the public's right to information, as established in the case-law of the Convention institutions. The relevant regulations must also provide for appropriate procedures, taking into account the technical aspects of the activity in question, for identifying shortcomings in the processes concerned and any errors committed by those responsible at different levels

evacuation. The Court also noted that to have the ability to inform subjects that danger was impending, there would have been need to establish temporary observation posts—however the state had been aware of this need as it had been informed by a surveillance agency, but also failed to act in this regard. In their judgement, the Court stated that, "[t]he authorities have thus failed to discharge the positive obligation to establish a legislative and administrative framework designed to provide effective deterrence against threats to the right to life as required by Article 2 of the Convention" (*Budayeva and Others v. Russia,* [2008]).

The Court additionally found that the state failed in its procedural requirement to protect the right to life by failing to investigate state responsibility through any judicial or administrative authority.

The case of *Budayeva* [2008] highlights a very important point mentioned in Chapter 1, that natural hazards lead to disasters when combined with vulnerability (Popovski, 2016, p. 97). By failing to reduce the vulnerability of its subjects, the state magnified the risk of the natural hazard, and exposed its subjects to human rights violations (Popovski, 2016). The state's negligence in ensuring its subjects safety from foreseeable forces rendered it directly responsible for the ensuing deaths, and perhaps at least partially responsible for the escalation of a natural hazard event into a natural disaster.

4.5.2 Slándáil-type Systems and the Right to Life

It now bears asking what relationship a system such as Slándáil could possibly have with the protection of the right to life? The doctrine of the margin of appreciation endorsed by the ECtHR suggests that states have latitude in their selection of mechanisms to protect the right to life, though in Court the adequacy of these mechanisms will be questioned. Fiduciary Theory itself demands that states proactively implement measures to protect life, and ensure that persons can live in secure and equal freedom, and not in fear of the devastating consequences of unchecked natural hazards.

Systems such as Slándáil function to provide emergency managers with information pertaining to the events unfolding in the aftermath of a disaster and provide them with information based on subjects' needs and infrastructural damage. The examples offered earlier demonstrate the raw potential of such systems to contribute to decision-making that results in the protection of life. The adoption of such a system by itself however does not negate, or fully satisfy, the state's responsibility to establish a regulatory and

administrative framework to deter threats to the right to life. The fiduciary state, to fully discharge its obligations, has a wide range of responsibilities throughout multiple phases of disaster management and cannot shirk its obligation to mitigate the threats of natural hazards, or plan for appropriate actions to be taken in natural hazard events. A system such as Slándáil cannot substitute for improperly maintained infrastructure for instance—it cannot stop an improperly maintained dam from bursting. A state that deploys a system such as Slándáil but neglects its other responsibilities remains just as responsible for any death and destruction that follows a natural disaster. Use of the system may result in mitigating some losses, however the state which neglects other areas of disaster planning could still not fairly claim to have done all that it was reasonably expected to do. If one adds the Slándáil system to the *Budayeva* [2008] scenario for example, responding authorities may have had access to information that could help them respond more effectively to areas in particular distress, however it would likely have done little to mitigate the damage and death caused by the immediate impact of the mudslide.

Whilst such systems cannot be the centre of emergency planning and response, that is not to say that they are not potentially a very effective tool that can be deployed by a state that rules with *solicitude*. And while such technologies have the capacity to contribute to saving human life, are reliable and potentially affordable and a state has been made sufficiently aware of their existence, it would seem that a state has a responsibility to adopt such systems where they are feasible and have a level of assurance that they can save life without disproportionately impacting other human rights obligations. A state that has the capacity to improve its emergency management powers and better discharge its fiduciary duty towards its subjects really ought to adopt such measures where they are practical, and failure to do so could be deemed a form of negligence in itself. In reality, such systems would be very unlikely to displace any other measures of emergency management. Interview participants representing emergency management agencies agreed that such systems would be supplementary to other methods of intelligence gathering in emergency events, and recognised the possible limitations of such systems in establishing the full picture of events, and operating consistently—therefore there is no evident desire for such systems to replace any existing measures.

It might also be noted that such systems can be used to assist in the procedural aspect of protecting the right to life as practiced by the ECtHR where effective investigations are conducted. Systems such as SIGE in particular can record actions taken by emergency managers, and information drawn from social media can be retained to be used as evidence in any investigations based on emergency response. To this extent, such tools can be used to enhance accountability, a topic which will be discussed in greater detail in Chapter 8. Acts of human and technical error, or negligence, can plausibly be identified and any parties responsible for wrong-doing potentially identified and disciplined or prosecuted if necessary.

The use of such systems, which provide emergency managers with potentially previously unknown and enriched information, also places these emergency managers in a unique position of responsibility which arises from their enhanced knowledge and position of power. Emergency managers will be in receipt of a potentially constantly updating stream of information that compels action, they will need to investigate and confirm reports of imminent danger and may be responsible for any act of harm that falls upon individuals should they be in a position to act and fail to do so. The existence of information regarding threats to life (perhaps an imminent flood risk identified by Twitter users) also places them in a position of needing to investigate and confirm these risks, and disseminate that information as effectively and quickly as possible in order to ensure that the wider public are made aware of risks to their lives. As information intensive resources then, systems such as Slándáil enhance decision-making with implications for human life, but also this knowledge that they generate places emergency managers as *enhanced knowers* in a position of being increasingly responsible for what they *do* with this knowledge.

In summary, systems such as Slándáil cannot be relied upon as the sole measure taken by the state as a means to protect the right to life but can be a useful additional resource that supplements other measures. Because such systems are information driven, they have the capacity to log and store information and act as records of decision-making, and evidence that can assist in any investigations or inquiries into emergency response efforts, potentially assisting efforts of ensuring the accountability of state actors for decisions made. As a system that enhances the knowledge of events of emergency managers, it also places the state in a position of increased responsibility to act and to share potentially life-saving information.

4.6 Conclusion

In this chapter, it was demonstrated that the Slándáil EMIS (and such similar systems) prove to be examples of artificial agents delegated with morally loaded tasks that, whilst ineffective in a vacuum, have the potential to improve the interactions between human agents and can harness the power of distributed morality to escape the tragedy of the Good Will. They can readjust the power asymmetry between emergency response agencies and disaster survivors and therefore empower both to function towards the same goal as part of a self-repairing multi-agent system, eliminating sources of entropy arising from natural evil.

The Slándáil EMIS (and therefore similar systems) can play a useful role in the protection of the right to life, however not on its own. Such systems can supplement pre-existing measures in emergency planning and response but the mere adoption of one measure does not indemnify a state of responsibility—the state's fiduciary duties require a range of measures to protect life, particularly in mitigation of natural hazards and not just response.

The system, and potentially others, can assist states in meeting procedural requirements of protecting the right to life by acting as repositories of information into investigations and inquiries into decisions made during disaster response, and could potentially help lead to the prosecution of negligent officials and thereby deter future negligent, life-threatening, actions.

Social media powered EMIS grant emergency response agencies more knowledge, and make them increasingly responsible to the extent that they have the power to act on this knowledge.

5 PRIVACY

5.1 Introduction

The second value to be examined in this thesis is that of privacy. Considering how rich in personal information social media feeds have a tendency to be, this is a value that will require extensive analysis from the perspective of the impact of Slándáil and similar systems.

Upon broadly outlining system functionality that has implications for the flow of personal information, an adequate theory of privacy compatible with the ethical framework of Information Ethics, and suitable to address the complexities of the information life cycle in the 21st century, will be unpacked. It will be argued that Helen Nissenbaum's theory of Contextual Integrity of Information (CI), supported by Floridi's Ontological Theory of Informational Privacy, is adequate for the task of analysis. Following this, the heuristic of CI will be applied to the context in which Slándáil is intended to operate.

Upon completing ethical analysis, the human rights implications of Slándáil vis-à-vis privacy will be analysed. Privacy, as understood in human rights law and as justified by Fiduciary Theory will be examined before analysis of the implications of Slándáil-type systems for this right.

Bearing in mind the transnational flows of personal information endemic in such systems, it will be necessary to outline a theory of extra-territorial application of human rights before analysis can continue. It will be argued that a gestalt model is the best approach to analysing states' obligations towards persons not located within its territory, which will then be used to assess the extra-territorial obligations of state's using systems such as Slándáil, which have the capacity to interfere with the rights of these people.

5.2 Features of the System with Implications for Privacy

Before a deeper exploration of the system's implications for the value of privacy, it is instructive to note features of the system that intuitively may have an impact on the privacy of individuals involved in natural disaster management information collection activities.

Social media sites are rich in personal information; any cursory search of publicly viewable Twitter feeds for instance can reveal much about a person, from their age, gender, and location to their favourite hobbies, favourite foods, political affiliations and more. The user generated content of social media sites can range from the mundane to the intimate, people can share anything from their personal stories, photography of landscapes to nude self-portraits (Lasén and Gómez-Cruz, 2009). In describing social media and personal identity, Floridi (2014, p. 63) remarks, "[n]othing is too small, irrelevant, or indeed private to be left untold". The sharing of such information is an inherent aspect of such services, which are platforms for communication on any topic, and promote interactions between friends and strangers unrestricted by geographical space and to an extent, time. The Slándáil system enhances the visibility of messages that are transmitted on such social media sites,[62] messages that have the potential to contain personal and potentially sensitive information not specifically intended to be shared beyond certain audiences. The system displays contents of messages to emergency managers, and can visualise the precise location of the social media users on a map.

In addition to enhanced visibility/accessibility of social media messages for emergency managers, the system also provides emergency managers with a platform to save and store this information. Saving and storing information acquired from social media sources is an option and potentially a necessity for emergency managers who may need to record their rationale for making a decision in an emergency situation.[63]

Beyond uses by emergency managers, the system will also store social media messages collected during an event for the purposes of machine learning and improving the system's efficacy in natural disaster situations, as well as review of decisions made by emergency managers. As illustrated in the dataflow diagram in the previous chapter, the system will extract terminology from social media messages in order to train the system to better recognise and filter relevant messages. The same logic applies not only to text content, but images also.

[62] At least, at present, where such messages relate to natural disaster.

[63] Both interviewed emergency managers noted the requirement of making record of information leading to decisions in an emergency context for the purposes of public enquires. Elaborating further, one stipulated that command and control logs were required to be recorded by law, and that information leading to a decision would potentially be recorded including images and social media message contents, though not necessarily fully reproduced messages but messages in a "resume" or abbreviated format and that such information would be stored in accordance with data protection law.

Two initial implications for privacy then, which will need to be analysed in greater depth in the following sections, are enhanced accessibility by potentially unexpected agents and temporally indeterminate storage of social media message content by public and private agents.

5.3 Privacy and Ethics

5.3.1 Towards a Theory of Privacy

Defining privacy can be an elusive endeavour; it has inspired a rich history of rigorous theorising on its content and substance, and generated contention between scholars as to what it is or should be (Tavani, 2008; DeCew, 2015). In its earliest conceptions, privacy was viewed as something of a boundary that separated the personal from the public, it was representative of a right against unnecessary government interference with the personal life, and was also uniquely associated with property (DeCew, 2015). It is a concept that has often been conflated with others, and sometimes contained no unique substance of its own as a result (Tavani, 2007).

The concept of privacy has evolved and grown more complex as scholarly work and technological contexts continue to advance in a world being re-ontologised by ICTs that change the nature of moral problems arising from the enhanced information life-cycle. By 1890, the moral challenges to the distinction between a private and public life posed by evolving ICTs[64] were beginning to emerge perhaps more seriously and resulted in more intense interrogations of the concept (Warren and Brandeis, 1890; DeCew, 2015). By 1890—with the advent of photography combined with the increased availability of affordable media through which information could be widely disseminated—work on elucidating the content of privacy arguably began properly when it was recognised that the unfettered transmission of personal information could amount to an infringement of privacy, which was argued to be "the right to be let alone" (Warren and Brandeis, 1890, p. 195; DeCew, 2015).

Without pondering the full gamut of privacy theories, it is useful to refer to Tavani (2007) for a very brief overview of older theories which may not be useful for analysis here because of the conceptual weaknesses that they embody. Tavani breaks privacy theories into two categories (which can be subdivided into two sub-categories), *Nonintrusion and Seclusion,* and *Control and Limitation*.

[64] That is, the question of appropriate domains for information relating to the person.

"[T]he right to be let alone" typifies nonintrusion theories of privacy, and has been criticised by Tavani (2007, p. 5) for its conceptual muddiness. These theories, Tavani (2007, p. 5) argues, conflate privacy as a descriptive condition, and a right, as well as conflating privacy with liberty.[65] As regards seclusion, this view of privacy holds that privacy is solitude, being inaccessible to others or ultimately, as perfect privacy is equated with perfect solitude, being "alone" (Tavani, 2007, p. 5). Such theories are pre-occupied with the notion of physical access (Tavani, 2007, p. 6). Whilst providing a descriptive account of privacy, unlike the nonintrusion views, the seclusion view conflates privacy with solitude (Tavani, 2007, p. 6). Requiring perfect seclusion is also a problematic view of privacy, one can have privacy outside of environments that are completely shielded from others (Tavani, 2007). Both interpretations of privacy are also pre-occupied with physical access, which may well render them obsolete in the technologically advanced infosphere where intrusion can occur without physical access to individual's space but access to publicly accessible records (Tavani, 2007; Nissenbaum, 2009). To that extent, such interpretations of privacy may be inapplicable to a range of 21st century issues where privacy intrusion can occur remotely, without breaches of physical boundaries.

Under the control interpretation of privacy, one has privacy when one has control over information relating to themselves (Tavani, 2007, p. 7), or essentially control over the information life cycle as it relates to the individual. This interpretation does not conflate privacy with other the other listed states or values, however it is unclear on the degree of control over one's information which one should have, and what kind of personal information one should have control over (Tavani, 2007, p. 7). Some personal information is nearly impossible to control—public personal information (PPI)—information including perhaps gender and race, information often available immediately to third party observers, from the contents of your shopping basket to whether or not you are at home (which your neighbours can easily determine)(Floridi, 2005; Tavani, 2007; Nissenbaum, 2009). Tavani (2007, pp. 7-8) argues that the distinction privacy interpretations should make between public personal information and private personal information (medical records for example) is unclear and such a theory is implausible if it grounds privacy as control over both of these information types. Tavani (2007, p. 8) expresses further scepticism of the control theory in that some control theorists argue that privacy can still exist when one has voluntarily disclosed all of their personal

[65] Liberty is not privacy, but the realisation of liberty is contingent on it to some degree.

information, because they chose and consented to do so—in this case the control interpretation conflates privacy with yet another value, autonomy.

In the limitation interpretation of privacy, privacy is said to be preserved when access to one's information is "...limited or restricted in certain contexts" (Tavani, 2007, p. 9). This interpretation of privacy holds that privacy exists when other persons have restricted (or even no) access to one's personal information (Tavani, 2007, p. 9). Tavani (2007, p.9) notes that this interpretation conflates privacy with secrecy in holding that one only has privacy to the extent that access to their personal information is limited and where a situation of perfect privacy is one where no information is held by one person about another.

These interpretations of privacy each have weaknesses amply demonstrated by Tavani (2007), each conflating privacy with other values or states which are closely related to privacy but contingent on it, and do not define its substance. They are conceptually weak and not normatively implementable, and in the case of the nonintrusion and seclusion interpretations in particular, excessively wed to the notion of privacy as a concept hinging on physical access, where intrusion is comparable to trespass or theft (Floridi, 2005). In the current iteration of the infosphere—where modern ICTs augment and enhance users abilities and re-ontologise the nature of our moral problems and mankind in general—marrying the concept of privacy to notions of physical access and physical boundaries is insufficient.

Humans transcend conventional physical boundaries and geographical borders as their personal information flows globally, with various PPI broadly accessible across different sources (social media for instance) and private personal information sitting on hard drives in various contexts (hospitals and insurance firms, statutory agency facilities, potentially even those belonging to states to which the person is neither citizen nor resident).[66] Defining privacy in this new world requires conceptual rigour. It needs to be

[66] Magnani (2005, pp. 13-14) eloquently illustrates the nature of man embedded in the infosphere, the modern day human:

> At present identity has to be considered in a broad sense: the externally stored amount of data, information, images, and texts that concern us as individuals is enormous. This storage of information creates for each person a kind of external "data shadow" that, together with the biological body, forms a "cyborg" of both flesh and electronic data that identifies us or potentially identifies us. I contend that this complex new "information being" depicts new ontologies that in turn involves new moral problems. We can no longer apply old moral rules and old-fashioned arguments to beings that are

defined as both a condition and right, and distinguish between circumstances where privacy is acceptably lost and where it has been violated (Tavani, 2007).

Tavani (2007)(and particularly James Moor, 1990, 1997) make an excellent effort in bringing forward privacy theory into the needs of the 21st century, and propose the theory of Restricted Access/Limited Control (RALC) to make up for the failings of preceding theories and approaches. This theory combines the stronger elements of the control and limitation interpretation whilst adding several vital elements. According to Tavani (2007, p. 30, citing Moor,1997, p. 30), under RALC, "...an individual has privacy in a **situation** with regard to others if in that situation the individual is protected from intrusion, interference, and information access by others'". The definition of "situation" is left ambiguous by design, but can refer to different contexts such as relationships; the storage, access, and manipulation of personal information on computers, or activities in locations (Tavani, 2007, p. 10). Privacy can fall into two categories, naturally private situations, where access and intrusion are prevented by physical boundaries, and normative privacy, where privacy is governed by the norms of the situation such as in the priest/confessor relationship (Tavani, 2007, p. 10). Tavani argues that in the former case, there are no norms and therefore privacy can be lost but not violated; however in the latter case it can be both lost and violated (Tavani, 2007, p. 10).

The RALC theory argues that lost privacy can be essentially accepted, personal information can and must in certain cases be disclosed, such as to medical professionals where it will then reside on computers (Tavani, 2007, p. 11). This information is protected by normative zones however, and access to it is restricted to only those with a valid claim to receiving it, and that persons should have some control over who has access, with disclosures and uses of personal information requiring elements of principles of choice, consent, or correction depending on the situation (Tavani, 2007, pp. 11–12). An important aspect of RALC is a principle of publicity, that is, the "...conditions governing private situations should be clear and known to the persons affected by them", so that they can exercise consent in an informed manner in a given situation (Tavani, 2007, pp. 16–17). The information life-cycle then should be regulated by contexts and norms. So long as personal information is transmitted between persons in

---

at the same time biological (concrete) and virtual, situated in a three-dimensional local space but potentially "globally omnipresent" as information packets. For instance, where we are located cybernetically is no longer simple to define, and the increase in telepresence technologies will further affect this point.

appropriate contexts, privacy can be lost but not violated, however when the normative rules of the situation are broken, privacy has been violated.

The RALC framework is a bold step forward for privacy protection that makes some adjustment for the nuances of the information-life cycle in the 21st century and the extent of control persons should expect to have over their personal information. However, without attempting to dismantle the entire theory, it has a somewhat flawed foundation in adopting the assertion that there are situations of natural privacy that are normless, and privacy can be lost but not violated. The examples offered are natural boundaries that prevent intrusion and interference, such as hiking or camping in the woodlands, situations whereby whilst one's privacy is shielded by obstruction, they are present in a potentially public location, and it would be eminently unfair to claim that it could not be violated, nor that there are no norms applicable to the situation (Tavani, 2007). This would appear to be a gap that leaves persons vulnerable to violation without any claim to call it that. It is difficult to call many situations naturally private with the augmenting and enhancing capabilities of modern ICTs, which can penetrate or circumvent physical obstruction, or on a more basic level, what is to stop some unknown party from stumbling into a naturally private situation?

The concept of the "naturally private situations" might cause more problems than it solves in its ambiguity, and this ambiguity can cause persons to be vulnerable to intrusion without a clear conceptual path to describing that they have been wronged. It is a concept which seems to owe something to interpretations of privacy bound to notions of physical boundaries that is quickly becoming moot in a world of CCTV, drones, satellites and thermal imagery. When does naturally private cease being such, and who is to know whether they are truly being shielded from observation?

Whilst Tavani (2007, p.15) acknowledges the necessity of protecting personal information that can be acquired from publicly available sources and extends the RALC framework to protect such information, the ambiguities of the naturally private still leave a troubling gap.

In the following, Floridi's Ontological Theory of Informational Privacy, and Helen Nissenbaum's theory of Contextual Integrity of Information (CI) will be examined to provide a more robust theory of privacy from which to work.

5.3.2 The Ontological Theory of Informational and Privacy Contextual Integrity of Information

Floridi's (2005) ontological theory of informational privacy situates the concept of privacy firmly within the framework of IE by arguing that human beings are informational in nature, that their information does not so much belong to them as it *constitutes* them as an informational entity. Before delving into the theory in depth, it is instructive to first briefly review Floridi's position on personal identity, in order to understand how the construction of the self and privacy are co-dependent.

Floridi (2014, p. 60) argues that the humans are defined by three selves that must be in harmony for humans to flourish, the *personal identity* ("who we are"), *self-conception* ("who we think we are") and *social selves* (essentially who others think and say we are). ICTs then have a great impact on these three selves, where services such as social networking sites allow us to create narratives that define us, or have narratives built by others that define us too; both of which, the self-conception and social self, feed back into personal identity—who we think we are and who others say we are feed back into who we are or can become (Floridi, 2014).[67]

Modern ICTs, Floridi (2014) argues, give us great freedom in defining ourselves, by presenting ourselves however we please, but simultaneously reduce our freedom in constructing our identities. Just as we can generate data that becomes information that defines us, information can emerge from other sources beyond our immediate control that can define us too.[68]

---

[67] On describing the self today, Floridi (2014, p.69) argues:

> The self is seen as a complex informational system, made of consciousness activities, memories, or narratives. From such a perspective, you are your information. And since ICTs can deeply affect such informational patterns, they are indeed powerful technologies of the self...

[68] Floridi (2014, p. 63) argues that:

> Any data point can contribute to the description of one's personal identity. And every bit of information may leave a momentary trace somewhere, including embarrassing pictures posted by a schoolmate years ago, which will disappear, of course, like everything else on the planet, but just more slowly than our former selves will.

And (Floridi, 2014b, p. 72):

> Recorded memories tend to freeze and reinforce the nature of their subject. The more memories we accumulate and externalize, the more narrative constraints are provided

If we are our information, then that which relates to us, which defines us and allows us to be defined (for better or worse), is very precious—it makes us who we are, and what we can hope to be. As information entities then, when disparate pieces of our information are scattered across the globe in different locations, we are transnational entities that are highly vulnerable to interference.

Floridi (2006, p. 111) views the right to privacy as something that shields the personal identity. Privacy interference, when the self is viewed informationally, is neither trespass nor theft, but a kind of assault or kidnapping and a violation of the integrity of the self—information does not belong to someone in the sense of physical possession, but in the sense of "constitutive belonging" (Floridi, 2005, 2006, 2013, p. 245). He argues that no physical removal of information takes place, nor is a person removed from their physical space, but the observed (or a piece of them)—to the extent that they are their information—is moved to the observer's space of observation (Floridi, 2005, 2006). For this reason, it supports a right to privacy in public, for intrusion or violation can occur even where no violation of physical space has occurred (Floridi, 2005, 2006).

By Floridi's (2005, 2006) account of privacy, privacy is a function of ontological friction, or obstacles to the transmission or accessibility of information in the infosphere. To illustrate this, a person in glass house for instance would have little ontological friction inhibiting the flow of their personal information; the more ontological friction there is in an environment, the less is known about someone, the less ontological friction there is in an environment, the more is known about someone (Floridi, 2005). Ontological friction then is the forces that constrain accessibility of information, and can arguably be either physical, normative or artificial (imagine the encryption of personal data).

Floridi's interpretation does not conflate privacy with other values, though it is strongly linked with values such as autonomy and self-determination (or more generally, liberty). It is not a reductive theory however, and it recognises privacy as contextual; privacy can be lost or violated. Floridi (2006) recognises for instance that it is not non-negotiable, and can be to some extent exchanged for other interests (he uses biometrics for security as an example). Because of the importance of privacy to the integrity of the self, neither does the theory propose normless situations where privacy cannot be violated.

---

for the construction and development of our personal identities. Increasing our memories also means decreasing the degree of freedom we might enjoy in redefining ourselves. Forgetting is part of the process of self-construction.

The theory is something of an accessibility theory of privacy, whereby if a lot information about someone is accessible they have little privacy—however it is not absolutist and does not state that accessibility is tantamount to violation. It offers guidance as to where privacy has been lost, but does not strictly identify where it has been violated.

The theory is normatively minimalistic and unacceptable on its own for the use for analysis; it does not provide a definitive heuristic for identifying privacy violations. It is descriptively and justificatory useful, but fails to describe what privacy as a right should entail in detail.

This theory can however supplemented where it leaves gaps—therefore the next task will be outlining and justifying Helen Nissenbaum's CI, which offers substance that can fill in the gaps left by the Ontological Theory of Privacy without contradicting it. By using the two in conjunction, we can justify the necessity of privacy for the person as an informational entity, and better understand where violations of it have occurred.

Helen Nissenbaum's (2009) research on privacy theory is built upon an effort to explore and determine the concept of the paradox of "privacy in public", to what extent privacy can exist in public spaces and the normative framework that could be implemented to preserve it. Nissenbaum (2009, p. 217) emphatically rejects the public/private dichotomy, and forcefully argues against the notion that information obtainable in public or from public sources is "up for grabs."

The issue of privacy in public is a challenging one in a world of technologically advancing ICTs and raises questions of expectations of privacy in public spaces, and the negotiation of boundaries—in the hyperhistorical society, information can be generated, disseminated and made accessible in an instant. Ampáro Lasén and Edgar Gómez-Cruz (2009, p. 207) share an insight into the dilemma of privacy in public in this hyperhistorical age that bears discussion here as it highlights the challenges of what could be deemed persistent digital photography.[69] Ampáro Lasén and Edgar Gómez-Cruz (2009, p. 207) discuss in depth the story of a colourful young woman with pink hair who captured the interest of Spanish Flickr[70] users after appearing in multiple shared photos on the group "Madrid"—they develop a sense of familiarity with her and post images of her, all without her knowledge and generated in public locations. The girl is later

[70] Flickr is a photography-based social media website.

approached by a member of the group and informed about it; she reacts with discomfort (Lasén and Gómez-Cruz, 2009, p. 207). The reported exchange encourages introspection among the group members, revealing divisions and questions relating to expectations of privacy in a public place (Lasén and Gómez-Cruz, 2009, p. 207).

This anecdote is quite emblematic of the challenges of privacy, and expectations of privacy in public. In the case above, the girl with pink hair certainly lost privacy, and, without consent, her information was abducted and brought into the space of the observer as Floridi argues. Her social-self took on a life of its own, and through the inferences of others she acquired data beyond her control that could potentially shape her identity. Not only that, but we can assume that through repeated photography of a person such as the girl with pink hair, much could be learned about her and her habits, from frequently visited locations to perhaps even her workplace—the long time exposure of her details could even place her in danger should someone become obsessed and seek to do her harm. Nissenbaum's CI was arguably designed to protect exactly such people as the girl with the pink hair from the kind of interference described above.

Nissenbaum's (2009) work is driven by an acknowledgment of the impact of near ubiquitous ICTs and their potential impacts up on privacy, of their ability to merge together facets of information about people from disparate sources (which may have been reasonably disclosed in particular contexts) and compile extensive profiles or otherwise yield new types of information about someone.

For Nissenbaum (2009, p.140-147), privacy is arguably of extrinsic value; she does not conflate it with any other values or states but emphasises its importance in realising other values including self-determination, freedom from informational harm and injustice, the preservation of relationships, and others. What privacy is as a right, Nissenbaum (2009, p.127) argues, is the right to the appropriate flow of information about oneself. She has created a thorough heuristic for identifying privacy intrusive practices in her framework of CI—this is a heuristic against which to measure new practices with privacy implications, in order to identify points of public discomfort, and, as it identifies inappropriate flows of personal information, can pinpoint precise aspects of new practices that violate privacy. Privacy then is a state of one's personal information flowing appropriately, and a right to an appropriate flow of this information (Nissenbaum, 2009).

In Nissenbaum's (2009) framework of CI, the appropriate flow of personal information is determined by the specific context relative informational norms; it is a theory designed such that information produced or ceded in one context does not enter another one with a different set of parameters (at least without substantial justification). Context relative informational norms include four parameters; contexts, actors, attributes, and transmission principles (Nissenbaum, 2009, p. 140).[71]

Using this theory, one can now argue more persuasively why the girl with the pink hair can claim violation of her privacy, though it remains a complicated task. Taking photos of persons and sharing them online is arguably a norm, and one can certainly expect it when they leave the privacy of their homes—the principal of reciprocity is active here, one may takes photos of you and you in turn are capable of doing the same. The girl with the pink hair's case is more extreme, however. The principal of reciprocity has been invalidated by what can only be described as collective, out-sourced stalking, where the

[71] The context is essentially the "situation" that Tavani (2006) refers to, though at a more descriptive level it can be defined as "...the condition of application, or the circumstances in which the act is prescribed for a subject" (Nissenbaum, 2009, p. 141). Actors are "...senders of information, recipients of information, and information subjects," (Nissenbaum, 2009, p. 141). Attributes are information types; that is, what the information pertains to (Nissenbaum, 2009, p. 143). Certain types of actors will be justified in requesting certain information attributes as per the norms governing a context, whilst others in a different context would have no legitimate claim to be imparted with certain attributes, as Nissenbaum (2009, p. 143) clearly illustrates with the following example:

> Informational norms render certain attributes appropriate or inappropriate in certain contexts, under certain conditions. For example, norms determine it appropriate for physicians in healthcare context to query their patients on the condition of their bodies, but in the workplace context for the boss to do the same thing would usually be inappropriate (an exception could be made for circumstances such as a coach of a professional football team enquiring about a player's heart condition).

Finally, transmission principles are "...constraint[s] on the flow (distribution, dissemination, transmission) of information from party to party in a context" (Nissenbaum, 2009, p. 143). The transmission principles equate to the "terms and conditions" by which information flows, or should flow, appropriately (Nissenbaum, 2009, p. 143). Examples of transmission principles include confidentiality, reciprocity, dessert, entitlement, and need (Nissenbaum, 2009, p. 143). Information transmission may or may not require consent based on the particularities of a context—transmission principles, unique to a given context, will determine whether information is given voluntarily and consensually or is simply coerced from a subject (Nissenbaum, 2009, p. 143). For example, when one is setting up a direct debit to pay their electricity bill, they voluntarily supply their bank details to the service provider and permit them to withdraw the billed amount every payment period. On the other hand, if one has become subject of a criminal investigation because they are suspected of embezzling funds from their place of employment, they can expect perhaps far ranging intrusion and can almost certainly expect their bank account details and transaction history to be acquired and accessed by the investigating authorities without their permission.

girl has been repeatedly photographed and become a subject of intense interest and discussion. There is a power asymmetry between her and the forum members; they are in a privileged position of access to her information that she could unlikely ever replicate with the potentially anonymous users of the forum—there is scant opportunity for reciprocity. The girl did not consent to this, and when notified she emphatically rejected the behaviour of the Flicker users. The girl with the pink hair, due to the inappropriate flow of her personal information, was essentially abducted in her capacity as an informational entity and placed under a microscope. In both cases, a strong argument for privacy violation can be made, and not just a case for innocuous and acceptable privacy loss.

CI provides more structure to the normatively minimalistic Ontological Theory of Informational Privacy. Ontological friction may determine the level of privacy enjoyed by someone, and ontological friction (constraints or obstacles to transmission of information) can be used synonymously with transmission principles. Whilst the ontological theory of informational privacy argues that privacy is a function of ontological friction, and one has more or less based on their information's accessibility, CI can be used as a tool to highlight with more certainty where lost privacy is inappropriate and a violation—where departures from contextual informational norms occur.

CI is broadly compatible with IE, and in fact the particular parameters of context relative informational norms can be transposed onto the model of a moral action as illustrated in Figure 3 in Chapter 2 with some symmetry. The context itself is synonymous with the moral situation; the actors are synonymous with agent and patients; attributes are embedded in the information process and agent/patient; and transmission principles cut through the shell, the factual information concerning the situation, and the envelope in how they deem the use of that information appropriate.[72]

As a final point, returning to the analogy of abduction; abduction is only considered so when one is removed from a location or placed in another without their consent, or without legitimate reason (consider illegal detention by state authorities). When one moves from one location to another voluntarily, it is not abduction; when one goes to work at risk of being fired for not attending, it is not abduction although there is a

[72] The moral actor, when dealing with information pertaining to another subject for instance, will act upon principles located within the shell—their morals—based on external information that they have—about the situation, the law, and ethical and social norms.

reasonable level of coercion involved; when one is arrested and detained on suspicion of committing a crime it is not abduction though it is done involuntarily and without consent. If we are our information, our information moves through the world with much the same constraints our bodies do, as determined by social, legal, and ethical norms. What privacy is ontologically and as a right is the *appropriate* flow of our information. Whilst a large degree of freedom is necessary to protect and shape our personal identities, and to protect us from informational harm, total freedom is implausible if not undesirable in a healthy society, a view going back to and endorsed by the original contractarians for instance; who recognised that "[m]an is born free, and is everywhere in chains"(Rousseau, 1974, p. 8). Some reasonable boundaries to our freedom are not an affront to our dignity, but are often required for the protection of the dignity at all.

At this stage an implementable theory of privacy rooted into the chosen framework of IE has been outlined and justified. Now, the privacy implications of systems such as Slándáil can be more fruitfully analysed.

5.3.3 Slándáil and Privacy

At this point, to comprehensively assess the impact of systems such as Slándáil on the value of privacy, it is necessary to implement CI as a decision heuristic. This requires comparing and contrasting the entrenched practice to the novel one, and entails several steps, outlined by Nissenbaum (2009, pp. 149–150) as:

- establishing the prevailing context
- establishing key actors
- determining affected attributes
- establishing changes in transmission principles
- red flagging where a departure from the normal occurs

Beyond this, where a norm departure is found, evaluation is required in order to determine the broader moral harms of the practice, and its impacts in values and goals within specific contexts (Nissenbaum, 2009, p. 182). This is not a task that can be completed in sum in this chapter, however the following chapters will individually address implications on values beyond privacy; to that end the CI and disclosive approach pursue the same goals and complement each other quite effectively.

The first point to establish is the traditional prevailing context. The context is that of emergency management and communication in natural disaster. The actors (or

agents/patients) in this context are variable, though the context is primarily characterised by agents within the emergency management agency,[73] and natural disaster survivors.

The typical method of communication between disaster survivor and emergency management agency remains the phone (though social media is, as argued, a burgeoning source of information). The disaster survivor can provide primary information, or eye-witness reports, by contacting emergency management agencies using an emergency phone number. Information attributes will be variable, and dependent on the particular circumstances of a natural disaster; however persons affected by disaster are likely to report infrastructure damage, and/or personal need of rescue or rescue of someone else. Information exchanges may entail personal information, such as names and addresses of the caller or details (including name and addresses) of persons who may perhaps be reported as missing or in danger in the event of a disaster. Attributes then in emergency communications from disaster survivors are likely to be environmental or personal.

Active transmission principles include reciprocity (emergency management agencies have a duty to inform the public with information regarding the natural disaster and response efforts), entitlement and dessert (the public are both entitled to and deserving of information pertaining to the disaster and response efforts as is the emergency management agency who require information to operate effectively), and though disaster survivors are not coerced into to making reports, the emergency management agency will have need of information reported and all relevant attributes that can assist them in emergency response, and are justified in receiving this information where it relates to the reporter or someone who is being reported on. It bears noting that reporting is likely to occur at least shortly before, during, and in the aftermath of a natural disaster event.

With the introduction of the new practice (data collection and processing by systems such as Slándáil), the arrangement is altered significantly. The context now changes somewhat due to overlap between emergency communications and response in natural

[73] Including emergency managers in command, and emergency response agents deployed to the field. There are other agents that play roles within the context of emergency management and communication in natural disaster generally, including scientists, meteorologists, and task forces. The level of abstraction adopted here will be limited mostly to directly consider the relationship between agents of emergency response and agents/patients involved directly in the natural disaster (survivors).

disaster, and the overall context of social media—therefore the moral situation is the combination of two contexts. The actors involved are the same for the most part, the emergency management agency including managers and responders, and natural disaster survivors. Additions are made however, including tentatively general social media users (persons can post a message about an event regardless of location or involvement in that event),[74] potentially private companies whose AAs will be processing the social media messages, and the AA, or the artefacts to which it can be broken down (Slándáil and its components).

Information attributes will largely be consistent with those shared in the traditional scenario, in both of which a large variety of different information types can possibly be shared. It bears noting firstly however the ability of Slándáil and similar systems to display the geo-location of where tweets originate from means extensive recording of persons' locations to potentially highly accurate degrees—there arguably is a superficial attribute change through automated visualisation of exact location of disaster communications on an interactive map, though this is an ambiguous case and no further argument will be made as it may not warrant further review to the extent that it does not provide ongoing tracking of particular individuals. Secondly, Slándáil also filters relevant images, and gives emergency managers access to visual information that may be limited[75] in the traditional scenario, and image content can be attribute rich. The traditional emergency manager/disaster survivor relationship is changed by enhanced image filtering from potentially civilian sources, however attributes may not necessarily reveal more or differ extensively from information received from a traditional phone call.

Because of the nature of social media, and because communications would not necessarily be targeted at emergency managers, there is also the possibility that the system will display information that is not relevant to emergency management. Therefore messages rich in attributes unrelated to an emergency manager's goals could be presented to them and saved to their systems, though the risk of this is entirely dependent on the efficacy of such systems in correctly filtering relevant messages.

[74] This can arguably be the case in the traditional scenario too, where it comes to persons outside of the location of a disaster event reporting missing persons, for example—though this category of actor bears inclusion as new simply because anyone can talk about an event and contribute towards the irrelevant noise whilst still potentially having their message processed.

[75] But not nonexistent when one considers the potential for CCTV cameras in public spaces, aerial photography and to a somewhat less intrusive extent due to small scale, satellite imagery such as from COPERNICUS.

Because this new context is a combination of two contexts (emergency communication and response to natural disasters, and social media), it is worth examining the typical actors and attributes in social media also. Social media consists of many actors, as already described (people and organisations), and information transmission can be multi-directional. Information attributes are highly variable and almost unlimited. Senders of information and information recipients will vary by service provider and privacy settings,[76] however if a post is made public, the information is accessible to anyone[77] with an account or account compatible with the given social media service. Twitter, for example, is a largely open platform that encourages interactions between perfect strangers. As a final note, not all information subjects appearing on social media are necessarily social media users. Social media users can post text, image, or video content about third parties if they please—the ethics of this must be judged on a case-by-case basis and in light of the particular context, and cannot be analysed in too much depth here beyond the examples already given of the girl with pink hair.

To address the transmission principles of the combined context: reciprocity is to an extent present in both contexts, reciprocity is present in both social media on its own and in natural disaster communications. In both cases information is expected to flow bi-directionally. Those who publicly share messages on Twitter (for example) openly do so with the understanding that it is publicly viewable and it can be engaged with and in return they can view and comment on the messages of others[78]—even where this is not explicitly understood or consented to, it is a demonstrated norm that defines the nature of the service. The emergency services and other statutory agents, for their part, and as demonstrated in Chapter 1, engage with this transmission principle and convey information, and interact with Twitter users on that platform.[79] Emergency management agencies' acquisition and processing of information from this source also operates on something of a more coercive basis including entitlement and need in the event of natural disaster, consent to process attributes is not strictly necessary when life

[76] Facebook for instance has granular privacy settings where the audience of a message can be explicitly chosen.

[77] Or anyone at all with an internet connection, and is potentially open to comment from anyone with an account.

[78] Though note that Twitter users with private accounts can still view public feeds.

[79] For an immediate example, the Twitter account of An Garda Siochana (https://twitter.com/GardaTraffic?ref_src=twsrc%5Egoogle%7Ctwcamp%5Eserp%7Ctwgr%5Eauthor) and The Police Service Northern Ireland (https://twitter.com/PoliceServiceNI?ref_src=twsrc%5Egoogle%7Ctwcamp%5Eserp%7Ctwgr%5Eauthor) which serve as good examples of this.

and death are at stake and the data to be processed is essential for the fulfilment of their duty to protect life and property. It must be noted however that principles of entitlement and need cannot be invoked where information is collected and processed from outside of the impact zone of a natural disaster, and where that is so the new practice can be flagged—and it is difficult to justify this practice, it will be returned to in more detail presently.

The potential presence of private companies or non-emergency response agents as intermediaries between the public and emergency management agency adds another layer requiring analysis. In the case of Slándáil at time of writing, numerous organisations including universities and private companies are involved to the extent that they are actively working on methods for the effective processing of social media messages in disaster events. The fine-tuning of such methods entails the collection and storage of social media content (including text and images) in order to train the system. Such an activity is one conducted contrary to the principle of reciprocity, and is done without consent of users who are also not notified of the potential uses of the content that they have generated. It can be difficult to ascribe this to entitlement or dessert when they are not statutory agencies that typically hold legitimate authority to implement coercive policies. The information in this sense is moved from one context into another, scientific research, and this is, at least *on a prima facie basis*, inappropriate.

Emergency managers are also capable of storing information obtained from social media, and as stated earlier, may be required to do so in order to provide rationale for decisions made in emergency response in any public inquiry. Saving social media messages locally is a norm departure (at least text content); although sites such as Twitter do have a "favourite" functionality to allow users to save and catalogue messages, though in the Twitter environment such messages will disappear if the user deletes them—the norm is essentially cloud storage based on ongoing user consent, and in principle is governed to some extent by reciprocity (to favourite a message a user must have a user account). The agents here, being statutory agents with normatively exceptional powers above those of civilians, can justify this through entitlement and need. The practice of saving documentary evidence is also functionally not dissimilar from the recording of emergency phone calls. If the entire system were to be located on site at emergency management headquarters, this may to some degree mitigate the

necessity of involvement of private and non-emergency management organisations, as suggested by one interview participant involved in the software development.

The first most obvious departure from norms in this analysis is the introduction of new agents that process information, the Slándáil system (and its components) and non-emergency management organisations including private companies. The saving and storage of messages from social media sites from such agents would also arguably constitute a deviation from the transmission principles.

Having identified norm departures, a red flag may be raised and this new practice may be considered, *prima facie*, a violation of privacy. This does not mean that the practice should be abandoned, it merely invites evaluation and justification for stepping away from entrenched norms.

A strong case can be made for the adoption of this new practice, and part of this case was made in Chapter 4. No definitive answer can be offered without the further analysis to be conducted in the remainder of this chapter as well as the chapters that follow. However it should be clear that systems such as Slándáil can make a valuable contribution to emergency management through providing enhanced situational awareness and through supporting decision making in potentially life-threatening circumstances. These systems have the capacity to contribute towards decisions that save life by revealing additional information about a moral situation that emergency managers may not have in their absence. They empower disaster survivors to contribute to disaster response in a simple but valid way (though it might be pointed out that this contribution may not even necessarily be their *intent*). They can accommodate collective action towards the removal of sources of entropy in the infosphere. Tentatively, the possibility that Slándáil and such systems can add to the efficacy of emergency response in natural disaster militates in favour of the adoption of such new practices. As to the existence of private actors in these contexts—as they have been delegated some surrogate statutory authority through their provision of a service to the emergency management agency, their norm departing inclusion in this context can be justified through the extension of this authority—so long as the processing, collection and

storage of social media messages is done exclusively to advance the capabilities of the system and messages are not retained when they have fulfilled a legitimate purpose.[80]

5.3.4 Case Study: The Intrusion Index and Technological Solutions for Promoting the Protection of Privacy

It is pertinent at this time to discuss the capacity of artificial agents (AAs) to assist in the preservation of privacy. Options for privacy preservation are diverse, however one such option is represented by the Slándáil system's use of a computational method called the Intrusion Index (II).

The II, like the SMM, uses natural language processing (in addition to named entity recognition) to identify the presence of nouns in social media messages, that is, the names of people, places, events, and organisations. Such data, particularly where all categories are present, can reveal much about a message sender or subject, and can be tantamount to lost privacy. The II logs the occurrence frequency of these named entities over time and represents the data on a graph, the peak of which (as seen in Figure 18), illustrates events or times of particularly intense occurrence of what is likely to be personal information.

In addition to representing such data on a graph, the II can also be used to visualise geo-located tweets (where the geo-coordinates are available, or where the location has been automatically inferred) containing named entities (see Figure 17).

[80] The private actors should not engage in the monetisation of social media messages and content.

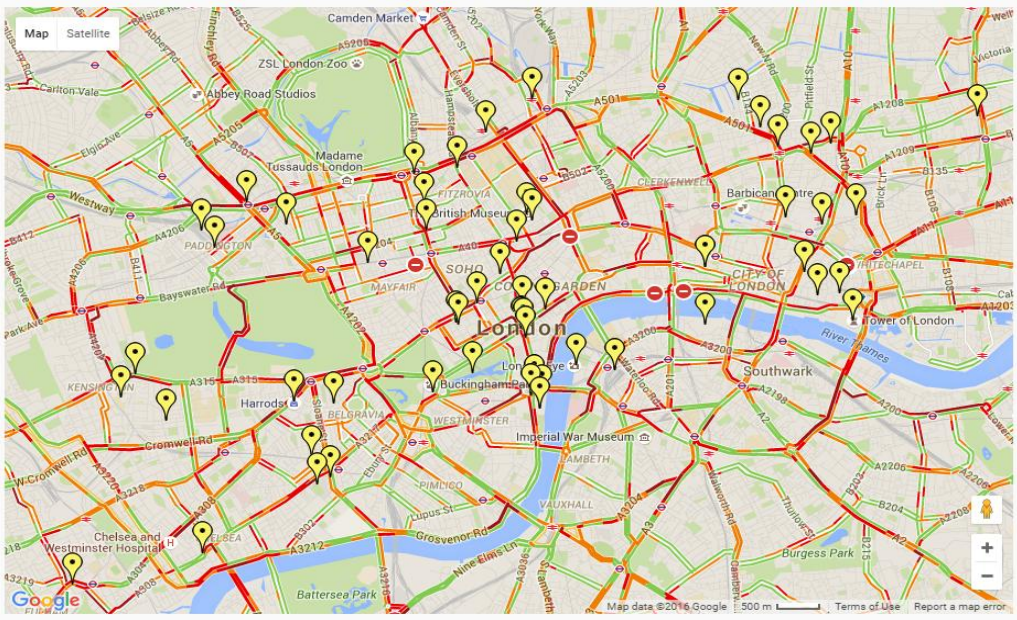

Figure 17: Geo-located Messages Containing Named Entities (Source: Slándáil-TCD, 2016)

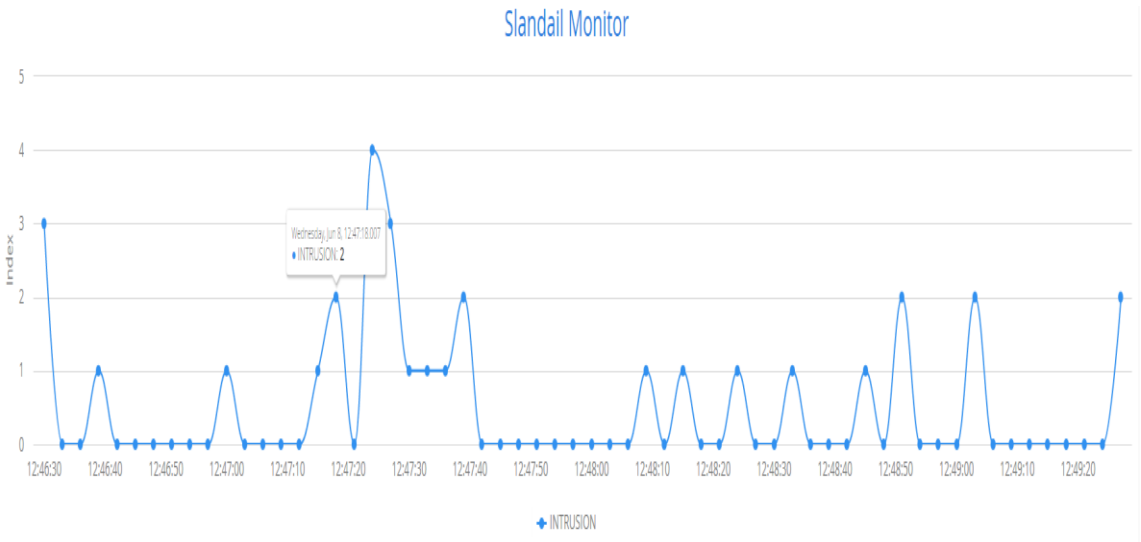

Figure 18: Graph of Frequency of Occurrence of Named Entities (Source: Slándáil-TCD, 2016)

The II is an AA that is inscribed with the "pro-ethical" condition of privacy in order to assist emergency managers in making judgements that can aid in the appropriate flow of personal information (Turilli and Floridi, 2009)—or put more simply, it is designed as a

moral enabler[81] that supports decision makers in the appropriate handling of personal information.

The II facilitates transparency in two senses. Firstly, it uses representation to make a problem (intrusion or privacy loss) more transparent (just as the example of costofwar.com does) and signals that action is potentially required in order to solve this problem, or prevent it from being exacerbated (perhaps the progression of privacy loss to violation as context shifts) (Magnani and Bardone, 2008). The II uses visual representation in the form of graphs and pins on an interactive map. To this extent, the Index serves to make emergency managers aware that they are potentially in possession of, or are likely to come into possession of, personal information, and therefore ought to deal with this in a manner most suitable to the particular requirements (or context relative informational norms) of the situation. Borrowing an analogy from Alison Adam (2005), it functions similarly to the alarm in a car that beeps continuously until the driver or passenger has buckled in their seatbelt. The Index then is advisory; it reminds the emergency manager of their duties towards protection of privacy without dictating or limiting action (or access) in dynamic circumstances.

Secondly, the Index supports transparency in the sense of disclosure (Turilli and Floridi, 2009). The quantitative output of the system can be disclosed to the public. When this disclosure is made, in conjunction with other information pertaining to the system including its functionality and the rules of governance of the system, the public can be fully informed about the system and its impacts. In this scenario, information is provided on how the system produces its information (a level of disclosure advocated by Turilli and Floridi, 2009, p.111). Public debate on the impacts of such systems is important, as the ethics of these systems may be redundant if they are met with mass disapproval regardless of adherence to ethical principles. With public dialogue, supported by disclosure, consensus can be reached regarding the implementation of the system, and the regulations determining its use and how personal information is dealt with. Reflection on ethics is an ongoing task, and it is only fair that the general public, whose interests are at stake, is given the opportunity to contribute to this debate which can result in additional insights into the relevant issues and perhaps an evolution of rules that govern the use of the system or its basic functionality.

[81] It does execute a moral action, but can support moral outcomes.

The II is just one example of an AA that has been delegated a morally loaded task with the intention of reducing moral inertia in a network, and increasing moral resilience. There are other options for the design and implementation of AAs that further shield privacy and act as agents of ontological friction. An AA could be designed for instance to anonymise content of messages, or perhaps blur facial features in images just as in Google Maps. There is a delicate balance to maintain however between the autonomy of emergency managers in discharging their duties to protect life, and the privacy of social media users or other information subjects. Reducing the information available to emergency managers may run the risk of endangering disaster survivors if inadequate information pertaining to their situation cannot be discerned. Anonymisation in the natural disaster response context might be counter-productive and unnecessary in a moral situation where perhaps low degrees of ontological friction are expected and norms support a relatively open flow of personal information.

Nonetheless, AAs delegated with morally loaded tasks can help achieve moral outcomes, not only in the scenario illustrated in the preceding chapter but in other situations. How they function and what their goal is, however, requires deliberation.

5.4 Privacy and Human Rights

5.4.1 Privacy and Fiduciary Duty

If the legitimate sovereign authority is required through its fiduciary duty to provide a regime of secure and equal freedom, under the rule of law, and provide conditions of non-domination and non-instrumentalisation in so doing, then the value of privacy in discharging this duty is evident. When privacy is violated, subjects of the state's authority can be easily dominated by the will of others (whether represented by state or non-state actors). Examples of this can be manifold: violations of privacy can result in subjects (and politicians) being threatened with blackmail by public or private actors. As another example, privacy in the voting booth can help prevent the coercion or retribution of others and the manipulation of elections. Privacy is a right that safeguards liberty, and as already suggested, a range of other values and rights, and can actively protect a functional democracy.[82]

[82] On the latter point, UN Special Rapporteur on the Promotion and Protection of Human Rights and Fundamental Freedoms while Countering Terrorism (Scheinin, 2009, p. 13), succinctly put it:

In addition to safeguarding subjects from domination, when Luciano Floridi's Ontological Theory of Informational Privacy is taken into account, privacy as a right takes on a dual importance—if one's personal information is being abused and used towards particular unjustifiable purposes, they are not simply being subjected to domination, but as informational entities are arguably being subjected to instrumentalisation as well—core parts of their being could be used in such a way that their dignity, and humanity, is denied.

5.4.2 Privacy and International Human Rights Law

In international law, the right to privacy has been enshrined in the *United Nations Declaration of Human Rights (UDHR)* (1948), Article 12 and The *International Covenant on Civil and Political Rights (ICCPR)* (1966), Article 17. The *European Convention on Human Rights* (1950), Article 8, Right to Respect for Private and Family Life, enshrines the right to privacy.

It is useful again to turn to some of the case law of the ECtHR in order to illustrate how the right to privacy (particularly informational privacy) has been protected in practice, and the rationale behind judgements made in such cases.

There are several instructive cases to consider that illustrate the Court's approach to the right to privacy, particularly as regards data or personal information usage. A particularly salient case is that of *Peck v. the United Kingdom* [2003], which dealt with the inappropriate dissemination by local authorities of photographic images captured by CCTV of the applicant committing an act of self harm in a public location. Here the Court decided there was a violation of Article 8, finding that inadequate efforts were made to minimise the interference with the applicant's privacy (that is, inadequate efforts at anonymisation) and the applicant's consent was not sought—such a case emphasised the need for *necessity* of such measures that interfere with privacy; here there was no necessity for the disclosures made in the manner they were made (without consent or anonymisation).

The Court has also found in favour of principles of data correction and limited data retention, as highlighted in the cases of *M.M. v. The United Kingdom* [2013], and *Khelili*

> Privacy is necessary to create zones to allow individuals and groups to be able to think and develop ideas and relationships. Other rights such as freedom of expression, association, and movement all require privacy to be able to develop effectively.

v. Switzerland [2011]. The former case dealt with the (essentially) forced disclosure of a police record to the applicant's prospective employer. The applicant's record had been very old, and the right by which the state retained and disclosed data relating to it had not met the quality of law, or benefited from adequate safeguards from abuse, therefore the Court found a violation of Article 8.[83] In the latter case, the Court established that the retention of inaccurate data was an Article 8 violation (the applicant's record had inaccurately referred to her as a prostitute). Whilst the Court agreed in principle here that data retention was justified in assisting with crime prevention with regard to the possibility that offenders could re-offend, here the use of the term was not based on sufficient evidence and was unnecessary.

The Court has also notably decided upon the use of far reaching surveillance powers. In *Roman Zakharov v. Russia* [2015], the applicant complained of a covert system of mobile

[83] In M.M. v. the UK [2013] the Court acknowledged that storage of such information fell under the scope of Article 8, and reaffirmed that "[e]ven public information can fall within the scope of private life where it is systematically collected and stored in files held by the authorities...". The Court noted that the caution, having been stored on police files, was available for disclosure long after the event, even when the event had been long forgotten by anyone, and argued that "...as the conviction or caution itself recedes into the past, it becomes a part of the person's private life which must be respected." The Court noted that the applicant consented to the disclosure, however given the circumstances of vetting for the role, she was given no real choice. The Court took the data protection aspect of the case seriously and reiterated that:

> The Court considers it essential, in the context of the recording and communication of criminal record data as in telephone tapping, secret surveillance and covert intelligence-gathering, to have clear, detailed rules governing the scope and application of measures; as well as minimum safeguards concerning, *inter alia*, duration, storage, usage, access of third parties, procedures for preserving the integrity and confidentiality of data and procedures for their destruction, thus providing sufficient guarantees against the risk of abuse and arbitrariness...

The Court found a violation of Article 8, deciding that the disclosure was not in accordance with the law, based on many short-comings of the law and procedures determining the disclosure, and a lack of effective safeguards from abuse, including:

> ...[t]he absence of a clear legislative framework for the collection and storage of data, and the lack of clarity as to the scope, extent and restrictions of the common law powers of the police to retain and disclose caution data.

Additionally:

> It further refers to the absence of any mechanism for independent review of a decision to retain or disclose data... Finally, the Court notes the limited filtering arrangements in respect of disclosures made under the provisions of the 1997 Act: as regards mandatory disclosure under section 113A, no distinction is made on the basis of the nature of the offence, the disposal in the case, the time which has elapsed since the offence took place or the relevance of the data to the employment sought.

telephone communication interception. The applicant complained of the general existence of law that allowed such surveillance, not that he was specifically subject to it. The law "Order no. 70" required mobile telephone service providers to install equipment enabling access to communications of service users by Russian security services. Here, whilst accepting the legitimacy of surveillance generally, the Court noted that unchecked surveillance posed a threat to the very democracy it was intended to protect.[84] Order no. 70 failed to meet the quality of law, the Court found that there were inadequate safeguards against abuse; the laws and procedures were inadequate to protect individuals from arbitrary interference with their right to privacy and that there therefore had been an Article 8 violation.[85]

The foregoing demonstrates how informational privacy is protected in practice, that for a legal subject to enjoy a regime of secure and equal freedom, under all that entails, they must be granted reasonable autonomy and freedom from unnecessary interference by state agencies. The state is not entitled to unfettered access and control over personal information, and is subject to legal constraints prohibiting them from dominating its subjects through unjustified intrusion in the private life. The state cannot dominate one's personal identity or alter the fate of its legal subjects through indefinite retention of data, especially where it inaccurately describes the subject. Without the right to privacy there is no secure freedom, the Court recognised that untrammelled

[84] In *Roman Zakharov v. Russia* [2015] the Court accepted the utility and necessity of secret surveillance in principle but noted:

> In view of the risk that a system of secret surveillance set up to protect national security may undermine or even destroy democracy under the cloak of defending it, the Court must be satisfied that there are adequate and effective guarantees against abuse.

[85] In *Roman Zakharov v. Russia* [2015] the Court stated:

> In particular, the circumstances in which public authorities are empowered to resort to secret surveillance measures are not defined with sufficient clarity. Provisions on discontinuation of secret surveillance measures do not provide sufficient guarantees against arbitrary interference. The domestic law permits automatic storage of clearly irrelevant data and is not sufficiently clear as to the circumstances in which the intercept material will be stored and destroyed after the end of a trial. The authorisation procedures are not capable of ensuring that secret surveillance measures are ordered only when “necessary in a democratic society”. The supervision of interceptions, as it is currently organised, does not comply with the requirements of independence, powers and competence which are sufficient to exercise an effective and continuous control, public scrutiny and effectiveness in practice. The effectiveness of the remedies is undermined by the absence of notification at any point of interceptions, or adequate access to documents relating to interceptions.

state power over private communications in the form of secret surveillance runs the risk of destroying democracy itself. The state's legitimate claim to authority requires solicitude in the means and laws that it adopts that have implications for privacy.

Interferences with the right to privacy are acceptable where they are justifiable and sufficient safeguards from abuse are in place, and rights can be derogated from or limited as explained in Chapter 2, where these limitations and derogations are themselves necessary to ensure a regime of secure and equal freedom under the rule of law. In the next subsection, the implications for derogation and limitations of the right to privacy will be briefly assessed.

5.4.3 The Limits of Privacy

As discussed, rights may conflict or total adherence to the protection or respect of a right may at times be a constraint on the provision of a regime of secure and equal freedom under the rule of law.[86]

Two of the major human rights instruments quoted earlier offer no explicit clauses for limitations to the right to privacy: the *UDHR* (1948) and the *ICCPR* (1966)—the intent of the articles is clear however, using language such as "arbitrary" and "unlawful" as qualifiers. In explicit UN Covenant limitation clauses, reasons for limitations are usually expressed as being necessary include public order, public health, public morale, national security, public safety, the rights and freedoms of others, and the rights and reputations of others (UN Commission on Human Rights, 1984, pp. 4–5; Sommario, 2012, p. 326).

The *ECHR* (1950) makes explicit mention of a state's right to restrict privacy as necessary in stating under Article 8, paragraph 2.[87]

---

[86] Consider for instance the *ICCPR* (1966) Article 20 requirement of prohibition on advocacy of national, racial or religious hatred that incites hostility or violence versus its Article 19 enshrinement of freedom of expression. In such cases states may pass anti-incitement (or "hate speech") legislations that place restrictions on one right in order to properly discharge protections of another.

[87] The ECHR (1950), Article 8, paragraph 2, reads:

> There shall be no interference by a public authority with the exercise of this right except such as is in accordance with the law and is necessary in a democratic society in the interests of national security, public safety or the economic wellbeing of the country, for the prevention of disorder or crime, for the protection of health or morals, or for the protection of the rights and freedoms of others.

Perhaps one of the most useful sources of instruction on the limitation of rights is The UN Commission on Human Right's *The Siracusa Principles on the Limitation and Derogation Provisions in the International Covenant on Civil and Political Rights*, which offers a comprehensive list on the conditions for justified limitation and derogation.[88] This report (1984, pp.3-6) essentially outlines the principles that constrain limitations, which can be summarised in the following points:

- No limitations or reasons for applying them may be permitted other than those provided for by in the terms of the Covenant.
- The scope of limitation should not jeopardise the essence of the right.
- Limitations should be provided for in domestic legislation (that is, they should have a domestic legal basis).
- Limitations should be applied exclusively to the purpose for which they were prescribed.
- Limitations should be subject to challenge and effective remedy.
- Limitations should not discriminate against particular groups.
- Where a limitation is deemed "necessary" it is based on grounds justifying limitation; responding to pressing public or social need; pursuing a legitimate aim, and is necessary to achieve this aim.
- The state shall not use means that are any more restrictive than necessary to achieve the goals of the limitation.

In short, limitations to rights must be lawful, strictly necessary, have a legitimate aim, proportionate, subject to safeguards preventing abuse, and provide for effective remedy.

These principles have largely been observed in the case law of the ECtHR, and so it is useful to turn attention to some illustrative cases to parse out the practice of limitations in the context of informational privacy, where the Court found no violations of Article 8.

In contrast to *M.M. v. The United Kingdom* [2013], *Leander v. Sweden* [1987] concerns an applicant who was dismissed from a naval base following a disclosure of information by the police to the armed forces who conducted a background check. The law regulating the retention and disclosure of information here met the quality of law test, and the measures were found to be proportionate and to meet a legitimate aim

[88] This report was the product of the work of 31 legal experts hailing from a number of different states and convened in Siracusa, Sicily, in 1984 (UN Commission on Human Rights, 1984, pp. 5–6).

(national security), therefore no violation of Article 8 was found regardless of the interference with the right.[89]

And in contrast to *Roman Zakharov v. Russia* [2015]*, Uzun v. Germany* [2010] also dealt with the issue of surveillance, though found the measures interfering with privacy acceptable. The case dealt with the GPS surveillance of the applicant and an associate, suspects of crimes perpetrated by extremist groups. The Court held that systematic collection of private data by any individual authority constituted a privacy interference, even where it contained no sensitive information and was likely never consulted.[90] The Court determined that an interference under Article 8 had occurred, and proceeded to examine whether it had been in accordance with the law. The Court found that the law enabling such surveillance was sufficiently accessible and met the foreseeable criterion. It also found that the applicant was protected by sufficient safeguards; surveillance was subject to judicial review and there was the possibility that GPS data obtained could be excluded from trial.

In the preceding cases, the right to privacy was limited in order for the state to discharge its fiduciary duty of providing a regime of secure and equal freedom. The limitations did not represent deviant state behaviour that conflicted with its duty, but were necessary

[89] The Court had to determine whether the interference was in accordance with domestic law, and to that end whether the implications of that law were foreseeable to the applicant and whether the law was accessible; and also if it was *necessary* in a democratic society (*Leander v. Sweden,* [1987]).

As to accessibility, the law was available in the Swedish Official Journal. And as to forseeability, the Court determined that the discretion of the police to store particular types of information was sufficiently circumscribed and limited to the purpose of protecting national security, and that it provided sufficient detail as to conditions for communicating that information. On the basis of this and that the law was adequately publicised, the Court stated:

> Having regard to the foregoing, the Court finds that Swedish law gives citizens an adequate indication as to the scope and the manner of exercise of the discretion conferred on the responsible authorities to collect, record and release information under the personnel control system.

In establishing whether the measure was necessary in a democratic society, the Court noted that this requires that there be a pressing social need and the measure is proportionate to the legitimate aim being pursued. The Court accepted the State's margin of appreciation in its assessment of pressing social need, but needed to determine if adequate safeguards from abuse were present. The Court found adequate safeguards present, with the process of storage and transmission of personal information tightly controlled and under the scrutiny of an ombudsman and parliamentarians, as well as with the potential of appealing decisions.

[90] Importantly, the Court noted that "[p]rivate-life considerations may arise, however, once any systematic or permanent record comes into existence of such material from the public domain..." in reference to interferences with privacy in public space (*Uzun v. Germany,* [2010]).

steps to preserve the safety of all those subject to their power. In each case, the restrictions were implemented in order to protect the state security or apparatus of the state necessary to protect state security (*Leander* [1987]), or deter the commission crimes. The state can restrict rights where they might imperil its ability to discharge its duties, or come into conflict with other rights, subjecting its subjects to the possibility of domination or instrumentalisation. In most cases, privacy is limited legitimately where it is lawful, necessary, proportionate, and subject to appropriate safeguards from abuse. Interference cannot be arbitrary, the state may take action that interferes with rights, but cannot itself become a force of domination, terrorising its subjects with the prospect of unjustified intrusion into their private lives.

The lawful requirement of human rights limitations is an important point to note. The quality of law requirement, which demands laws be accessible and foreseeable, is consistent with Lon L. Fuller's internal morality of law, endorsed by Evan Fox-Decent as a requirement of the rule of law in the fiduciary relationship between fiduciary and legal subject (Fuller, 1977; Fox-Decent, 2011). Fuller essentially argued that law should be sufficiently general, publicised, consistent, plausibly followed, and clear and understandable (Fuller, 1977; Fox-Decent, 2011). The internal morality of law respects the agency and dignity of subjects, respecting them as morally responsible agents, purposive beings, that can be judged for their wrongs, that are not merely acted upon by the state for breaking rules (Fox-Decent, 2011). In this sense, it ensures that the law and legal system treat people as ends and not means, and prevents them from being dominated (Fox-Decent, 2011). Where a law is not accessible and foreseeable, the legal subject does not have their agency respected—they cannot reasonably be expected to comply with a law that they have no access to or understanding of its implications for their personal behaviour, they become objectified and exposed to arbitrary interference.

The other manner in which privacy can theoretically be limited is through acts of derogation in emergency, a topic covered well in Chapter 2, therefore the particularities of derogation in emergency will not be dwelt upon here—though it must be considered for its implications on human rights treaties without specific clauses. Criddle and Fox-Decent (2012) argue that where limitations clauses exist in treaties, derogations are unnecessary, or at least require special justification. For states party to the *ECHR*, it is intuitive that no derogation is strictly necessary to limit the right to privacy, as it has an

explicit clause. For the rest of the world, the *ICCPR* only offers ambiguity although it could be argued that either derogation or standard limitations are applicable. There is intent in usages of the word "unlawful" for instance. Scheinin (2009, p.8) for one, argues that Article 17 of the *ICCPR* is subject to the standard principles governing limitation, due to the wording of the text. This is reasonable; privacy has not obtained the status of a peremptory right, and it would be absurd to require derogation any time an interference is required in the public interest—that the *ICCPR* has no *explicit* limitations clause is arguably an oversight on the part of the drafters and cannot set a standard by which states can operate.

The application of measures suitable for addressing the exigencies of emergencies to normal times might be concerning, as highlighted by Gross and Ní Aolain (2006). It may well be that the temporary suspension of some norms exemplified by derogation is preferable to the permanent institution of norms designed for exceptional circumstances—such mechanisms are designed to protect democracy and are ultimately still subject to checks and balances; research has even shown that derogations are characteristic of states that respect the law, and that states do not typically derogate insincerely (Hafner-Burton, Helfer and Fariss, 2011, pp. 692–694).

The fact remains that the state which rules within the parameters of the fiduciary relationship, acknowledging the internal morality of law and human rights, is obliged to create law that respects the agency and freedom of its subjects, law that should be subject to checks and balances, and should be open to challenge. The state can plausibly institute law of a permanent character that is capable of addressing the exigencies of an emergency, and where it can do this in-keeping with the requirement of the internal morality of law, and where privacy interfering measures are necessary, legitimate to the aim pursued, proportionate, and subject to safeguards from abuse—such a practice is justifiable and not necessarily a cause for concern. Where normal law that does not subject citizens to undue interference is sufficient to address the needs of the emergency, and where human rights treaties provide explicit clauses, the state should use conventional limitations and not derogations. Where there is doubt however, for those not party to the *ECHR*, derogations are more liberally justified if not demanded by the situation—such states should be under extra scrutiny by the international community in emergency, something which international notification would prompt.

Even where a state is party to a treaty with explicit limitations clauses, however, if no means it has at its legal disposal are sufficient to address the exigencies of the situation,[91] it should derogate from the right nonetheless. If exceptional measures have no basis in domestic law and no derogation is made signalling the temporary nature of deviation from the norm, the state is fully accountable to its subjects and the international community. The means utilised, as required by fiduciary theory, should nonetheless be necessary, proportionate, and never violate peremptory norms.

5.4.4 Slándáil-type Systems and Privacy

The Slándáil system is not a system that empowers individual or group surveillance at its base, and it does not at a most basic level (and in its currently existing form) explicitly enable the intentional systematic collection of information pertaining directly to individuals for the compilation of dossiers or files. Neither should this be possible so long as the terminology database, or dictionaries, by which the system (and similar systems) operates is adapted to only filter messages relating to natural disasters.

With that being said, the system does enable emergency managers to store messages and content on their systems, which could be rife with personal information. In addition to this, messages processed by the system will be stored on databases potentially owned by private organisations or academic institutions—in providing a service with public ramifications to statutory agencies (emergency management agencies), the statutory agency must also accept responsibility for the actions of these private or academic institutions which have been folded into the process of disaster response in the provision of technology that requires their active involvement.

When emergency managers and supporting institutions come to obtain and store or communicate personal data, an interference with the right to privacy has occurred. Storage for both service providers and emergency management agencies will likely be necessary. As mentioned earlier, emergency managers may need to retain information that provides evidence of the rationale for decisions made, and service providers will need to collect and store information in order to train the system.

With regard to the storage and potential transmission of information,[92] an interference with the right to privacy arguably occurs. At this point the interference will require a

[91] That is, where the only measures it can take to address the needs of the situation are not reasonably provided for in domestic law.

[92] Say from emergency managers to the media, or between partners providing the service.

basis in the domestic law of the state experiencing an emergency, and probably within the territories where partner service providers are located (recall that Slándáil partners are located across four states). Beyond this, recourse to the use of the system and storage of information obtained by the system is dependent on necessity, proportionality, and adequate safeguards from abuse.

It should be noted that if there is a substantial basis in domestic law for the use of such systems and subsequent related data uses, despite being operational in an emergency context, the system can be deployed without an official declaration of emergency and derogation from the right to privacy. In the event that domestic law is insufficiently clear or ambiguous in its implications for the use of such systems (particularly in terms of foreseeability, though certainly the law should be accessible) for potential data subjects, and with regard to the severity of the emergency, a declaration of emergency by the executive would be required and international notification of derogation.

The onus is on the state deploying such systems to justify the necessity and proportionality of the use of such systems and the subsequent handling of personal data. The grounds for use of Slándáil and similar systems (and any law that authorises such use) can be justified on the grounds of pressing need, including public health and safety. They are valid tools in pursuit of a legitimate aim, which is ultimately saving of life and property, and can provide emergency management agencies with the information to help them discharge this (fiduciary) duty effectively.

In terms of proportionality, the question becomes more complex. Other activities undertaken with little to no human rights implications can be utilised in the disaster management cycle in order to mitigate the destruction of property and threat to life. However, as demonstrated in the last chapter, enhanced access to frontline information in disaster response can provide previously unobtainable information that can be used to save life. The measure—EMIS that harvest data from social media—may not be the least intrusive measure to safeguard human life but provide a definite potential to do so, and as such can be argued to be a proportionate measure (along with necessary data storage and transmission) to the aim sought. Systems such as Slándáil can provide emergency managers with enhanced situational awareness to support decisions in resource allocation, potentially supporting resource allocation to areas or persons in distress which the agency would potentially otherwise have been unaware of.

It would appear that the onus of justifying the interference in rights that systems such as Slándáil represent may in some ways necessitate interference. When data is recorded and stored for the purposes of providing evidence of the rationale for decision making in any review or inquiry, it stands to reason that such practice can also be used to justify the interference where, for instance, a tweet lead directly to saving someone's life, particularly where the threat to that person would have otherwise gone unknown to the emergency manager.

As to the collection and storage of social media messages by private actors—tacitly approved by emergency managers utilising and relying upon the service—this practice is arguably proportionate when the end result is that the improved dictionaries can be used to better identify quality information that can save life. Once the data has served its useful purpose of training the system, however, it should be deleted or its contents anonymised to the greatest extent possible without compromising the efficacy of the system. The data should not be monetised or shared indiscriminately, by either private actors or statutory agencies.

As to the proportionality of transmission of information acquired by the system, the transmission of information during or after the event may result in further interference (bear in mind the case of *Peck* [2003] in particular). For private actors acting in partnership with the emergency management agency, the communication of data between databases across jurisdictions may be another example of an interference. For the system to operate effectively and produce life-saving information, this may be necessary and proportionate. Where the entire infrastructure of a system is consolidated onto the emergency manager's premises this would be minimised and a moot point.

The transmission of information obtained from the system by the emergency management agency may also constitute an interference (recall *Peck* [2003] again). It could be argued that the onus of responsible information sharing is on the originator of the information, and once that is shared (on a public social media platform), what the emergency management agency does with it is irrelevant. On the contrary, the emergency management agency remains responsible for the appropriate use of any information it obtains. If it obtains information, including images, it is duty bound (out of respect for the dignity of the information subject) to transmit this only where appropriate and in an appropriately edited form, particularly with regard to the

attributes of the information. An emergency manager for instance could not justify the transmission of an uncensored image of a naked man in flood water to the media without justification, regardless of it already being public, for it would be complicit in increasing the audience of that potentially harmful information. Transmission of data on a need to know basis, and not in an arbitrary manner, is acceptable from the perspective of proportionality.

A further challenge for the condition of proportionality is that whilst information is collected and processed from within specified geographic limits, this specified area can be extended well past the particular area affected by disaster. In this case, broad and unnecessary interferences can occur whereby persons not on site of the disaster but whose messages are filtered due to containing relevant terms, have their messages and potentially locations made visible to emergency managers, and potentially collected and stored by the system service providers. On the one hand, this can be useful if an 'outsider' expresses concern for a relative located in the affected area that the emergency manager can follow up on. On the other hand, a bulk of irrelevant information is more likely to be processed. The mere possibility of persons using a social media platform to express concern is not a substantial enough reason to collect information from an area that substantially exceeds that experiencing a disaster, therefore this could be too indiscriminate to be proportionate. The specified geographic area where tweets are monitored needs to be generally symmetrical to the site of emergency or likely emergency, or as close as the system will allow, in order to satisfy the condition of proportionality.

In the light of this discussion, it may be concluded that the system and related or dependent data storage and transmission activities can be proportional considering the legitimate aims that they serve.

Apart from this, necessary safeguards should be in place that ensures that the acquisition, access to, storage, and transmission of personal information is appropriate and not arbitrary—this might entail procedures and regulations governing the usage of information, possibly involving independent review. Most emergency management agencies will already have protocols determining the appropriate transmission of information within and without the agency, and the introduction of systems such as Slándáil may not necessitate change.

A question arises as to when it is acceptable to activate systems such as Slándáil; should it be perpetually monitoring social media feeds for natural disaster related information, or should it only be activated upon confirmation of an emergency or imminent emergency? In the absence of an ongoing or imminent event, activation of the system cannot be said to be necessary on the grounds of pressing social or public need, and it could not be deemed proportional to the aims sought, which themselves come into question outside of a natural disaster. If derogation is required to legitimate the use of such a system, its ongoing activity is even more difficult to justify, as it would require the declaration of an indefinite emergency.

Wherever domestic law is not adequate to justify the usage of such systems, a state may resort to declaration of emergency and derogation from relevant articles of human rights treaties, on the basis that the threat is exceptional and constitutes a threat to the organised life of a community. This raises an interesting question as to the severity of an event that justifies activation. Where domestic law is sufficient to authorise the use of systems such as Slándáil, the system can arguably be activated to address needs arising from a natural hazard event that is in itself life threatening without perhaps being severe enough to invoke the status of 'disaster,' there may be minimal disruption caused by the natural hazard event but it may nonetheless pose an escalated danger (perhaps elevated water levels in a river with a harsh current) or threaten disaster. Such borderline cases can justify the activation of a system; it would enable authorities to monitor areas in real time where individuals may be at increased risk of drowning. If however domestic law did not authorise this, the emergency management agency may have to rely solely on traditional methods for detection of persons in danger—a declaration of emergency cannot be made if the danger is merely perceived, it must be *real and exceptional*.

5.5 Privacy and Territory

In its currently existing form, the Slándáil system can be used to collect social media messages pertaining to a crisis in any specified geographical territory. As well as this, at least the prototype system can operate in such a way that data is exchanged between databases belonging to partner service providers; therefore there is a transnational flow of information. The international flow of data in this regard, and the capacity to use such systems to capture information pertaining to events unfolding outside of the jurisdiction of emergency management agencies, raises the question of to whom human rights are

reasonably owed, the state's legal subjects or all those affected by the state's use of power? This is a complex issue, and one that will need to be outlined and parsed out before answers can be offered. The extra-territorial applicability of human rights will need to be determined before a proper analysis of the privacy implications for systems such as Slándáil for those located outside of the jurisdiction of end-user emergency managers can be properly undertaken.

5.5.1 The Problem of Extra-Territorial Application of Human Rights

The extent of a state's responsibility to either ensure or respect the human rights of persons outside of its physical borders is a point of debate. The use of conjunctive language in the *ICCPR* (1966) Article 2 is ambiguous enough as to leave it open to varying interpretations (italics added for emphasis) (King, 2009; Van Schaak, 2014, p. 27; Milanovic, 2015):

> Each State party to the present Covenant undertakes to *respect and to ensure to all individuals within its territory and subject to its jurisdiction* without distinction of any kind...

The *ECHR* (1950) offers less ambiguity on the same subject, but remains open to interpretation, in stating that:

> High Contracting Parties shall secure to everyone within their jurisdiction the rights and freedoms defined in Section 1 of this convention.

The ambiguities of these texts have led to three competing interpretations of States' extra-territorial human rights obligations; the *narrow* view, the *protective* view, and a *gestalt* model (King, 2009; Margulies, 2014; Van Schaak, 2014).

Under the *narrow view*, states owe positive and negative human rights obligations to persons located within their territory and jurisdiction (King, 2009; Margulies, 2014; Van Schaak, 2014). In order for these criteria to be satisfied, persons would have to be within the borders of the state's sovereign territory, and the area occupied by that person would have to be subject to the control of the state.

Under the *protective view*, the state owes duties to two classes of persons; those within its borders, and those subject to its jurisdiction (therefore human rights obligations would be owed even to persons within a state's territory even if it did not exercise control over that territory) (King, 2009; Margulies, 2014; Van Schaak, 2014).

In the third view, the emerging *gestalt model*, the state owes positive obligations to persons within its territory and jurisdiction, negative duties to respect the rights of persons "...without territorial limitation," and must ensure (that is, it holds positive obligations) the rights of a person where it has the capacity to do so and as defined by its relationship with that person (Margulies, 2014; Van Schaak, 2014, pp. 29, 48–49).

The narrow view in particular is concerning and would appear to be against the essence of the purpose of human rights; it is conducive the human rights abuse without the human rights abuser being subject to any accountability—in extremis, the population of potentially occupied territories or jurisdictions are owed no human rights obligations (positive or negative)(King, 2009). When the *narrow view* is adopted by states, particularly in an age where some states are empowered by advanced capabilities in communication interception that have transborder impact (Ball, 2013; Shubber, 2013), it is natural that the right to privacy is particularly at risk. The United States has in particular generally adopted the *narrow view* on its extraterritorial responsibilities (Milanovic, 2015), and has possibly one of the most advanced known electronic surveillance systems in the world (Ball, 2013).[93] The uncertainty existing around the extra-territorial application of human rights means that the *narrow view* can be invoked as a shield for states that might be overreaching in their interferences of the rights of others, it is exploited as an opportunity to indemnify them of any responsibility or accountability.

An example of a state invoking the *narrow view* as a shield is the ECtHR case of *Weber and Saravia v. Germany* [2006]. This case concerned the extra-territorial interception of telecommunications between a German journalist and an employee of Montevideo City Council by German authorities. In its defence, the state argued that the application was "...incompatible *ratione personae* with the provisions of the Convention..." as both applicants were resident in Uruguay (*Saravia v. Germany,* [2006]). In this particular case, the state essentially argued that it owed no human rights obligations to persons outside of its jurisdiction (or at least, the jurisdiction of the ECtHR). Rendering this perhaps an even more extreme invocation of the *narrow view*, was that it denied having liability in this case despite one applicant having a legal relationship with the state (a German

[93] The position of the US in this regard is not monolithic, and was perhaps relaxed under the administration of Barack Obama; however, at time of writing, under the Donald Trump the narrow view is evidently being doubled-down upon (Milanovic, 2015; Lomas, 2017; whitehouse.gov, 2017).

citizen). This application was found inadmissible, but is emblematic of the toxic potential of a narrow view that is dismissive of a state's ability to exert power over people outside of its territory and jurisdiction.

The case law of the ECtHR on extra-territorial human rights interferences as a whole is inconsistent and offers little definitive normative guidance. The ECtHR's decisions have ranged from showing deference to the *narrow view*, to accepting a *gestalt model*.

In numerous cases involving Turkey's support of the Turkish Republic of Northern Cyprus (TRNC) in Cyprus, including *Loizidu v. Turkey* [1996], *Manitaras and Others v. Turkey* [2008], and *Andreou v. Turkey* [2010], the Court affirmed that Turkey was responsible for ensuring and respecting human rights in the occupied territory (the territory was brought within the jurisdiction of Turkey through effective administrative and military control), and even, in the case of *Andreou v. Turkey* [2010], for bringing individuals within their jurisdiction by virtue of cause and effect (a bullet fired by an agent of the TRNC struck the applicant who stood outside of their territory).

In perhaps an even more liberal interpretation of a state's extra-territorial human rights obligations is the case of *Soering v. The United Kingdom* [1989]. Here the applicant faced murder charges and the possibility of extradition to the US, where he would face treatment tantamount to a violation of Article 3, prohibition of torture or inhuman or degrading treatment or punishment. In another cause and effect rationale, the Court found that the United Kingdom would be responsible for any such violation that the applicant faced upon extradition.

In one of its more conservative judgements however, deferring to a *narrow view*, was the case of *Bankovic v. Belgium and 16 Other Contracting Parties* [1999]. In this case the applicants complained of the North Atlantic Treaty Organization (NATO) bombing of a Serbian Radio and Television HQ that resulted in numerous deaths. As the site of occurrence of the bombing was not located within the jurisdiction of the contracting parties, the Court found the application inadmissible.

It might be noted that the *protective interpretation* is also not realistic, positing that a state has far ranging human rights obligations in situations that might be far removed from its capacity to ensure rights in a given context—it could be faced with the impossible task of ensuring human rights where it exercises no real power or tangible influence.

The uncertainty surrounding the extra-territorial application of human rights represents something of a lacuna in human rights theory and practice, and a gap that must be satisfactorily filled before questions of the extra-territorial implications of human rights interferences can reasonably be answered. The *gestalt model* is the best hope of justifying extra-territorial human rights obligations that are plausible for a state to adhere to. In the following subsection, Hugh King's (2009) tri-partite model of the extra-territorial application of human rights will be argued to be an appropriate model, with the support of fiduciary theory.

5.5.2 A Fiduciary Solution

Elsewhere, the researcher has unpacked and examined how fiduciary theory can support a case for a *gestalt model* based on Hugh King's tri-partite typology of jurisdiction (Hayes, 2017). Here, the argument will be repeated as it has important implications for the analysis of activities which have extra-territorial implication for human rights.

Under Fiduciary Theory, the state is "...responsible to its subjects alone for the provision of domestic legal order" (Fox-Decent, 2011, p. 109). Despite this, the state still has the capacity to exert irresistible discretionary power over persons not within its territorial borders. As the foregoing case law fully demonstrates, they can become subjects of its *de facto* sovereignty. The law authorises the state to provide a regime of secure and equal freedom for its legal subjects, from whom its power flows, but not for persons outside of its territorial borders, not for non-citizens in foreign countries or "strangers" as the case may be (Fox-Decent, 2011; Hayes, 2017). However, these "strangers" who are subject to a state's irresistible power remain human beings with dignity, and they too have the capacity to place the state wielding this power in a *de facto* manner under obligation, proscribing their instrumentalisation and domination (Fox-Decent, 2011).[94]

Though the state which exercises power over a stranger does so without the authorisation of the law of stranger's land, a fiduciary relationship is still triggered—the state assumes the position of *de facto* sovereign in its relationship with the stranger. The state cannot in these circumstances rule on behalf of the strangers, however it cannot

[94] Fox-Decent (2011, p. 109) argues that for these foreign subjects, "[t]he state's power remains irresistible and administrative in nature, and strangers too have an innate right of humanity capable of placing the state under obligation" and that "[a]rguably, our innate right of humanity alone requires the state to act subject to fiduciary constraints regardless of whether they are citizens or strangers."

subject them to a regime that dominates or instrumentalises them—at a minimum it must respect the rights of the stranger, that is, it has negative obligations (Criddle, 2014). Maximally, where a state wields near total control over strangers, positive obligations are entailed (Criddle, 2014). Under the reasoning of Fiduciary Theory, the extra-territorial human rights obligations of a state are essentially proportional to the degree of control it holds over the territory or persons to whom it subjects to its power. Because human rights obligations are determined by the nature of power and influence in the relationship between the state and stranger, Fiduciary Theory endorses a *gestalt model.* By Fiduciary Theory, the state holds positive and negative obligations within its own territorial borders, however when it exercises *de facto* sovereignty by subjecting strangers to its power, its jurisdiction is extended and the fiduciary relationship is activated, though the shape of it is quite different.

Hugh King (2009) offers a tri-partite typology of jurisdiction consistent with the *gestalt model* and which appreciates the relationship between State, legal subject, and stranger. King argues for three different categories of jurisdiction with different corresponding packages of rights obligations:

- territorial based jurisdiction
- jurisdiction based on non-territorial factors
- jurisdiction based on a factual relationship

In the case of *territorial based jurisdiction,* persons who fall within the lawful jurisdiction of the state (that is, within its territorial borders where it at least exercises total control) are owed both positive and negative human rights obligations (the state must respect and ensure their rights). States may occupy foreign territories, and where this is so their authority flows from international as opposed to domestic law (of the occupied)(Criddle, 2014). The state's authority is lawful but remains *de facto*.[95]

[95] As Criddle argues (2014, p.13):

> …international law entrusts the occupier, like an international trustee or mandatory power, with a guardianship responsibility to establish basic security and safeguard human rights for the benefit of those within occupied territory, including both its own forces and the local population. Thus, principles of trust and fidelity lie at the heart of the international law of occupation—despite the fact that there may be little actual trust, and perhaps even deep-seated enmity, between the occupier and the populace of an occupied territory. As with the mandate and trustee systems, a state that governs territory under belligerent

Here, the State's human rights obligations are circumscribed by legal competence and the extent of factual control, that is, human rights obligations are discharged in accordance with local law to the extent that this law is compatible with a regime of secure and equal freedom, and to the extent that the occupying power has the capacity to discharge its obligations, for its control may be tenuous (King, 2009; Criddle, 2014).

In the case of *jurisdiction based on non-territorial factors* a state has "...lawful competence based on non-territorial factors" (King, 2009, p. 548). The state's obligations are relative to its legal competence (King, 2009). Persons included in this category include nationals (legal subjects, or citizens) located abroad—the state's obligations are circumscribed by the national law of the state where that legal subject resides (King, 2009). In this scenario, the state still holds discretionary, administrative power over its citizen (to whom it may have to issue a passport to ensure freedom of movement, for instance (King, 2009))—however, its power is limited and it is not in a position to ensure all of the rights of its citizen.

The final category, *jurisdiction based on a factual relationship,* refers to where "...a state, through its agents, acts beyond its lawful competence, it brings any person affected by its act within its 'jurisdiction' for the purposes of the ICCPR and the ECHR" (King, 2009, p. 551). Here, the obligations of the state are commensurate with its level of control over the individual, though King (2009) notes that these will often be negative in nature. The state as fiduciary, acting outside of its lawful authority, is still required to respect the rights of the distant subject. The innate morality of the fiduciary relationship places the state under obligation to respect the dignity of the stranger. Though the state acts without legal authority, the fiduciary nature of the relationship means that it should be seen as a legal one, subject to legal principles—this relationship must be moderated by human rights, commensurate with the power exercised by the state (Hayes, 2017).

It now bears asking what the implications of this are for the right to privacy. To achieve an answer to this, it is fruitful to be cognisant of Floridi's PI, and in particular, the Ontological Theory of Informational Privacy. Where a state obtains personal information through some action (such as surveillance) about an individual that can be considered a foreign stranger located abroad (neither citizen nor resident), a *factual relationship* exists and the fiduciary relationship is activated (though circumscribed greatly). The

---

occupation serves as a temporary 'guardian' or 'trustee' and bears corresponding duties of loyalty and care to the people under occupation.

state possesses personal information, that at least partially constitutes this stranger's identity, and must treat it with solicitude and it must be afforded the same protection as would a citizen's. The stranger must be protected from domination or instrumentalisation from the misuse of this personal information. The state's grounds for even obtaining this information must also be based on the principles regulating limitations of rights and derogations in emergency, that is, a state cannot collect and use the personal information of strangers without good reason, and without adequate safeguards and the possibility of effective remedy where abuse of this information occurs. A part of this person will reside in this state, and as such this person must be treated on the principle of non-discrimination, and their data cannot be arbitrarily interfered with on the basis of not being a citizen—the fiduciary duty of the state requires the provision of secure and *equal* freedom to those subject to the state's power. Negative obligations are owed. To this extent, where metaphysically an aspect of the stranger is based on the territory of the state, there is some overlap with *territorial based jurisdiction.*[96]

Information de-territorialisation by way of the persistent transnational flows of data are a challenge to the old Westphalian model of sovereignty, and are forcing the necessity for greater collaboration and cooperation between all states in order to respect and ensure the rights of persons whose data flows throughout territorial borders.

5.5.3 Implications for Slándáil-type Systems

The Slándáil system has the capacity to be used by emergency managers in one territory to collect information from social media feeds of persons located behind the territorial borders of another state. This can occur where emergency managers specify a geographical region in another state's territory. There is also the possibility of passive collection of social media messages from a foreign territory, as the messages collected are dependent on what the service provider's API sends to the backend of the system—the probability of this is unknown but it remains a possibility. Additionally, where the infrastructure is not consolidated in its entirety on an emergency management

[96] It may (and often will be) that a stranger will volunteer information to a service provider (for example Twitter or Facebook) whose databases are located in the state. The premise is similar here, however this entails positive obligations. The State hosting the service providers that have access to this personal information are responsible for the protection of the privacy of the stranger. The State has a responsibility to enact and enforce laws that will deter the exploitation of their personal information that would be tantamount to instrumentalisation or domination.

premises, information transfer will be transnational in nature as it is exchanged between partner organisations (universities and private organisations).

Based on the foregoing analysis, the emergency management agency holds at least negative obligations regarding the right to privacy of all persons whose personal information they come into possession of. This exercise of power triggers a fiduciary relationship, and commensurate responsibilities are activated. This means that the emergency management agency is bound by the principles of limitation and derogation. They do not have a right to indiscriminately interfere with the privacy of persons outside of their territory by virtue of their geographical location, in fact, persons affected by their use of power are drawn into their jurisdiction and acquire the benefits that entails.

In practice, many uses of the system on a second territory outside the sovereign control of the state using the system will not be justifiable. It would fail to comply with necessity even where the second state is experiencing an emergency—the emergency management agency, presumably not present in that state, will not have the ability to act on the information it is receiving. It also fails to be proportional for this reason.

This does not mean that such a use of the system would be prohibited in all circumstances. Where cross-border emergency operations are active and the emergency management agencies are working in close collaboration, this use can be justifiable. Where a second state solicits the resources of the state utilising the system in an emergency, such use would also be justifiable, on the condition that it is provided for in the domestic law of both states, is necessary, proportionate and subject to safeguards from abuse, or a derogation is made.

As to the partners who are communicating data across borders; they must be viewed as being party to the emergency response efforts and become surrogates of state power, exacting it as contracted by the state experiencing an emergency. These partners are constrained by human rights in their usage of personal information, and the contracting state should be considered responsible for any unreasonable interference for which they might be responsible. Using the logic unpacked in subsection 6.5.2, the states hosting these partners are also responsible for ensuring respect for the right to privacy of all persons whose personal information is collected by the system and might be present on databases located in those states. Where the state utilising the system during emergency response cannot be assured that the domestic law of partner states

adequately protects the right to privacy, or if the states hosting that data are likely to arbitrarily demand access to such data from partners, deployment of such systems may entail privacy violation that multiple states could be implicated in, and its use should be prohibited until such a time as it can be guaranteed that no arbitrary interferences with the right to privacy will take place throughout the information pipeline as it passes through multiple territories and overlapping jurisdictions.

5.6 Conclusion

This chapter contributed to the overall disclosive analysis by identifying the privacy implications of Sláindáil-type systems from both an ethical and human rights perspective.

From the preceding analysis it is clear that Slándáil-type systems pose an innate threat to the value of privacy which can be realised through inappropriate use of such systems. Using CI in particular, it is apparent that aspects of the configuration and potential uses of such systems deviate from established norms; new actors (or agents) are introduced to the relevant contexts (AAs including Slándáil and its components) and potentially private industry partners; emergency managers may have access to information that they cannot justify possessing, which is not necessary for the commission of their duties (irrelevant information or information from outside the disaster impacted zone); emergency managers can indefinitely store archived information and possibly disseminate it in ways not appropriate to the context of emergency management. The system, under CI, is a *prima-facie* violation of privacy. This does not prohibit its use, it simply needs to be justified. Chapter 4 illustrated its potential to save life, and undertook some of this work. Beyond this, the use of such systems and data generated by them needs to be regulated such that personal information does not migrate into any further contexts (unnecessarily), and is used exclusively on an as-necessary basis.

A human rights analysis demonstrates that states can act within the limits of their authority when deploying such systems, though risk exceeding this authority where domestic law does not clearly provide for its use, or its use would not be strictly necessary, proportionate, or benefit from safeguards from abuse. Even where no domestic law authorises its use, derogations would serve as an alternative should the threat posed by natural hazards be sufficiently serious. Nevertheless, there are numerous opportunities for states to exceed the limits of their authority, through indefinite retention of personal information, unnecessary dissemination of personal

information, or the collection of personal information from outside of a disaster impact area (particularly where it originates from outside the territory of the State) as some examples. Use of the system presents many opportunities to fall foul of human rights obligations, and as such systems and their uses need to be vigilantly regulated.

The purpose of this chapter was to interrogate the implications of the system's design and use in an effort to understand its ethical and human rights risks. Only in so doing can one propose solutions that can mitigate these risks. With a fuller understanding of these potential challenges, Chapter 9 will return to the task of proposing potential solutions that can help the ethical design and usage of such systems.

6 JUSTICE

6.1 Introduction

In this chapter, the implications of Slándáil-type systems for the value of justice will be analysed through the lenses of IE, and Fiduciary Theory.

Here, following John Rawls (1958, p. 25), the concept of justice will be taken to mean broadly "...the virtue of practices where there are assumed to be competing interests and conflicting claims—persons will press their rights on each other," and that such conflict must be dealt with in a fair manner, as "[j]ustice can be conceptualised as fairness.... which includes fair distribution (distributive justice), fair and reliable procedures (procedural justice), [and] fair retribution for evil and good done (restorative justice)" (Mordini *et al.*, 2009, p. 210).

This chapter will begin by examining human vulnerability in disaster in order to establish who in society is most vulnerable to the impacts of disaster, and will then proceed to examine the digital divide in order to discern who this phenomenon primarily impacts, and if there is any overlap between those more vulnerable to disaster and the digitally excluded, arguing that indeed there is. It is important to outline from the outset who precisely may be victims of injustice in disaster scenarios and as modified by the introduction of social media powered EMIS in disaster/emergency management.

Subsequently, an IE based approach modified by Capability Theory and adopting a Prioritarian logic will be used to evaluate the potential impacts of Slándáil-type systems on the value of justice.

Finally, Fiduciary Theory will be used to analyse the human rights based implications of the system, with a particular focus on discrimination—discrimination being an integral aspect of injustice insofar as it refers to fairness.

6.2 Vulnerability, Disasters, and the Digital Divide

Before embarking on a thorough analysis of the ethical and human rights impacts of systems that harvest data from social media during emergency response to natural disaster as framed by the value of justice, it is instructive to examine and outline the overarching context of existing structural inequalities in society that render vulnerable populations more acutely disadvantaged upon impact and in the aftermath of these

disasters. As will be reviewed in this section, such vulnerabilities are marked by personal or group characteristics (which may intersect) such as race, class, age, ethnicity, gender, and disability, and which may be reinforced by prevailing social, political, economic, and cultural conditions.

In addition to unequal experiences and outcomes after natural disasters, potentially vulnerable populations differ in their access, proficiency and ultimately engagement with ICT services on the basis of ingrained structural inequality. This phenomenon, known as the digital divide, is the gulf between ICT users and non-users.

The first issue raised here, that of unequal experiences in natural disaster based on personal characteristics arising from structural inequalities in society, is of obvious concern to emergency managers who will be in control of and need to decide upon the allocation of scarce resources based on need. The second issue raises a more contemporary concern with direct relevance to the present research, that is, digital representation of vulnerable populations on the internet.

This section will demonstrate that vulnerable populations are prone to acutely negative outcomes in natural disasters in comparison to more privileged populations, and what is more, in times of crisis where timely and relevant information is important in directing decisions in disaster response and the allocation of resources, the experiences and needs of these populations are at risk of not being reflected by digital information resources such as social media.

6.2.1 Unequal Experiences of Natural Disaster

It has been argued that the consequences of natural disasters are not a "natural" phenomenon so much as the outcome of the social, cultural, economic and political environment that the affected find themselves in (Neumayer and Plümper, 2007, p. 552; Menon, 2010, p. 310).[97]

[97] Roshnir Menon (2010, p. 310) eloquently argues that natural disasters are:

> Less a single destructive event than a social process unfolding within a particular environment and social context, a large earthquake, volcanic eruption or flood can unearth the bare inequalities of social development, which places some people more than others at risk, while undermining their capacity to mitigate, survive, endure, or cope with the consequences of such catastrophe.

Due to the economic, political, and social make up of a society some will be more vulnerable to the impacts of disasters than others, which is to say stratifications in society lead to stratification of disaster experience. In the context of natural disasters, those who are "vulnerable" or lack "resilience" are those who lack the ability to anticipate, endure, and recover from the impacts of natural disasters (Fothergill, Maestas and Darlington, 1999; Masozera, Bailey and Kerchner, 2007; Neumayer and Plümper, 2007; Zack, 2016).

Citing an unpublished paper by Cutter *et al.* (2001), Masozera *et al.* (2007, p. 301) outline some of the population characteristics that influence vulnerability and the reasons why those characteristics influence vulnerability. These characteristics, along with the (paraphrased) brief descriptions provided by Masozera et al. (2007) are:

- **Socio-economic status:** Individuals who have wealth can absorb loss better, particularly with the assistance of insurance, safety nets and entitlement programmes.
- **Gender:** Family care responsibilities and low wages pose a challenge for women's recovery after disasters.
- **Race and Ethnicity:** Language and cultural barriers restrict access to post-disaster funding and can result in occupation of hazardous areas.
- **Age:** Age, from the very young child to older persons, can affect mobility.
- **Residential Property:** Expensive homes incur higher replacement costs, and mobile homes are more vulnerable to hazards.
- **Renters:** Renters typically have less financial resources, lack access to information about financial aid after disaster, and may lack shelter options after disaster.
- **Education:** Education is linked to socio-economic status, and lower education impacts capacity to understand both warning and recovery information.
- **Health status:** Pre-existing ill health can increase morbidity in disaster, and lack of access to health insurance can increase vulnerability to disasters.
- **Social dependence:** People who are dependent on social welfare services are already marginalised and require additional supports after disaster.
- **Special needs populations:** Persons such as the homeless can be invisible in recovery efforts. Though not specifically referenced in the work of Cutter *et al.* (2001), as cited by Masozera et al. (2007), this characteristic might also include

persons with physical or mental disability who may experience difficulties in evacuating or understanding warning information.

The characteristics listed here are not mutually exclusive and can (and do) intersect, which may well further increase disaster vulnerability and ultimately shape an individual's experiences in disaster. A poor, sick person living in a mobile home for instance will face a combination of disadvantages and will be heavily impacted by a natural disaster, making their experience potentially more egregious than perhaps a poor but healthy person that owns their own home. The above list should also not be taken as being exhaustive, and only provides a snapshot of factors contributing to vulnerability.

A wealth of disaster research focuses on differential experiences in disaster, and a corresponding wealth of evidence has been generated showing that the presence of these listed characteristics can have serious deleterious outcomes for the disaster affected, and in some cases examines why these differential outcomes are so. The following three subsections will highlight evidence supporting inequalities based along the lines of class/income, race and ethnicity, and gender.

6.2.1.1 Class and Income

Research by Masozera *et al.* (2007) provides some useful insight into the impacts of natural disasters on the poor, using Hurricane Katrina in New Orleans as a case study. These researchers found that the poor were less likely to own vehicles and were as a result disadvantaged in the response phase of emergency—they suggest that this may have contributed to the rather large number of people that subsequently sought shelter in the Superdome (20,000 to 30,000 people) (Masozera, Bailey and Kerchner, 2007, p. 303). Masozera *et al.* (2007, p.304) also found that the poor were much less likely to possess flooding insurance, thereby constraining their recovery even more. Examining the global exposure[98] of the poor to disasters, Namsuk Kim (2012, p.203, 208) found that the poor have been more exposed to disaster.[99] The poor are also twice as likely as non-

[98] Where exposure is defined as "...the probability of natural disasters being realised with measurable impacts," and using the following measurable aspects of exposure, "the number of potential disasters.... the number of people potentially killed in disasters; and... the number of people potentially affected by natural disasters,"(Kim, 2012, p. 197).

[99] The global population living on USD 2 per day experienced 121 disasters in the period 2000 to 2009 in comparison to 101 experienced by the population above that threshold in the same period (Kim, 2012, p.203, 208) . It might be noted that this measurement is a very conservative metric of poverty that is likely to ignore the experiences of those living in relative poverty in

poor to be affected by disaster (Kim, 2012, p. 203).[100] Kim (2012, p. 208) found that the poor tend to live in more disaster prone areas, arguing that "...poor people are exposed to natural disasters not only due to the increase in the probability of being hit by one, but also because of greater concentration in risky areas due to migration, higher-population growth, or less pro-poor growth."

6.2.1.2 Race and Ethnicity

Examining the disaster experience differences between race and class (once again using Hurricane Katrina as a case study), Elliot and Pais (2006) make numerous important findings. On evacuation timing, it was found that members of the Black population living outside the city of New Orleans were 1.5 times more likely than similar White persons to leave after rather than before the hurricane (Elliott and Pais, 2006, p. 308). Within the city, although finding that a small population of mostly African-Americans reported never leaving the city, it was found that income was a stronger predictor of evacuation timing—residents in the income range of USD 40,000-50,000 were twice as likely to evacuate before the hurricane than those at the level of USD 10,000-20,000 (Elliott and Pais, 2006, p. 308). Low-income Black people, rather than specifically Black, or low-income people, were the most likely to remain during the disaster (Elliott and Pais, 2006, p. 308). Poverty and a lack of transportation were among the reasons offered for this divergence (Elliott and Pais, 2006, p. 309).

Addressing experiences in recovery, Elliot and Pais (2006, pp. 309-310) found in addition to renters and boarders being less likely to return to their homes a month after the disaster, Black workers were 3.8 times more likely than White workers to have lost their pre-Katrina jobs.[101]

Further to this, in an extensive literature review of research that documents differences between race, ethnicity, and class, Fothergill, Maestas, and DeRouen (1999) trace numerous important divergences in disaster experience. In this review (centred largely on American literature and experiences), research was reported by stage of the disaster cycle. At the *preparedness* stage, the researchers reported that ethnic minorities (that is,

---

states with advanced economies. Interestingly, Kim (2012, p. 203) finds that the gap between poor and non-poor disaster fatalities is closing.

[100] in the early 2000s 50 percent of the global poor were affected by disaster in comparison to 27 percent of the non-poor (Kim, 2012, p. 203).

[101] Black workers with household incomes of USD 10,000-20,000 were twice as likely to have lost these jobs than Black workers on a household income of USD 40,000-50,000)(Elliot and Pais, pp. 309-310).

non-White persons) had been disadvantaged by preparedness information that was available only in English; were less likely to have had disaster preparedness education; and that they were less likely to have useful emergency items and preparedness at the household level (Fothergill, Maestas and Darlington, 1999, pp. 158–159).[102]

At the *physical impact* stage the researchers found that numerous studies report disproportionately high fatality and injury rates among ethnic minorities across a range of natural disasters—one reason offered for this discrepancy was the standard of accommodation occupied by ethnic minorities, which is often old and unreinforced (Fothergill, Maestas and Darlington, 1999, p. 161).

At the *emergency response* stage language barriers are evidently a problem; in America emergency response agencies had too few bilingual speakers to communicate with non-English speakers (in the context of the literature reviewed, Spanish and Asian disaster affected); English language radio services had superior and more accurate information than non-English services; and conceptual differences between languages lead to problems such as differing concepts of spatial relations, rendering locating homes on maps problematic for response agencies for instance (Fothergill, Maestas and Darlington, 1999, p. 163). There is also evidence of differential responses by relief personnel along racial or ethnic lines; research has provided examples of Black communities having their power restored only after it was restored in White communities, and having received less assistance by relief organisations generally; additionally media coverage has been found to focus predominately on majority White communities to the exclusion of communities with larger minority populations (Fothergill, Maestas and Darlington, 1999, pp. 163–164).

At the *recovery* stage, yet more inequalities are borne out. Minorities face challenges at the recovery stage due to having lower incomes and savings, higher unemployment, less insurance and less access to information (Fothergill, Maestas and Darlington, 1999, p. 164). During recovery, it was also found that cultural misunderstandings between recovery agencies and minorities can lead to the construction of houses not adapted to their needs (Fothergill, Maestas and Darlington, 1999, p. 165).

[102] Household preparedness was based on having Items including flashlights, battery-operated radios, food and water supplies, first-aid kits, latches and cupboards and having given earthquake instruction to children (Fothergill, Maestas and Darlington, 1999, p. 159).

6.2.1.3 Gender

Women in particular can be highly vulnerable to natural disasters and their impacts, a vulnerability that can be compounded with lower social status[103] and in societies where culture, economic systems and politics interact in a manner that supports particularly sharp structural inequalities between men and women.[104]

Menon (pp. 311–312) argues that the interaction between external shocks and the every-day vulnerabilities of women disadvantage women to a greater extent than men, and that vulnerable populations are not at risk simply from external shocks, but that their marginality in society renders their lives a "permanent emergency."

Research has borne out disparities between men and women in disasters, with female mortality rates climbing higher than men's after disasters in both developing states and those with advanced economies—consider that in the wake of the Indian Ocean Tsunami 40,000-45,000 more women than men perished,[105] and in the Kobe, Japan earthquake, 1.5 times more women than men perished (Neumayer and Plümper, 2007, p. 555; Menon, 2010, p. 321). Research by Eric Neumayer and Thomas Plümper (2007, p. 560) found that female life expectancy is more adversely affected by disasters than that of males' (a result which is moderated by higher levels of women's socio-economic rights).

[103] Though Menon (2010, p. 311) states that women and men at high socio-economic status suffer in approximately equal number (in terms of numbers of deaths recorded from the disaster), a result also reported by Neumayer and Plumper (Neumayer and Plümper, 2007, p. 552) .

[104] Elaine Enarson (2014, p. 39) eloquently argues that:

> Those who bear the burden of disasters are predominantly women—the very poor and landless, single mothers, home-based workers, those who live with (and care for) the chronically ill, marginalized women (indigenous sex-workers, trans-women), and those who live without men (widows, lesbians, women heading households)...

And highlighting the intersections of the female gender and other characteristics that can compound vulnerability, she adds:

> ...women and girls figure large among the frail elderly (predominantly female), the very poor and landless (predominantly female), the overworked (predominantly female), the poorly housed and illiterate (predominantly female), and reproductive health needs (predominantly female).

[105] The mortality rate of women compared to men ranged from 1.2 to 2.1 times more female than male deaths across Indonesia, Sri Lanka, and India (Menon, 2010, p. 321).

Probably compounding the problem; It has been argued that disaster management as a discipline and practice has tended to neglect the circumstances of women and reflect a male bias (Enarson, 2014, p. 38).[106]

The circumstances of women often render them more vulnerable to the impacts of disaster both before and after the event. In the case of the Indian Ocean Tsunami, for example, mortality differentials arose as a product of a "...lack of information about evacuation warnings and shelter options, culturally restricted mobility, and responsibilities within the family that obliged women to stay behind to care for children and the elderly" (Menon, 2010, p. 321). In the aftermath and during recovery stages of disasters, women's outcomes are adversely affected by exposure to violence and sexual exploitation where there is a breakdown of law and order, and they suffer from unequal allocation of resources where boys and men are favoured over women and girls (Weist, Mocellin and Motsisi, 1994; Neumayer and Plümper, 2007; Menon, 2010).

Women, often with socially ingrained reproductive roles and livelihood responsibilities so tied to the household, face extreme challenges in recovering their pre-disaster socio-economic status (Enarson, 2014). Enarson (2014, p. 41) argues that high pre-disaster female poverty levels normally increase, particularly among single mothers, and that this is a phenomenon observed across all states, rich and poor—Enarson (2014, p. 42) illustrates this with the example of Kobe, Japan, where single mothers faced higher levels of post-disaster unemployment and had greater difficulty in accessing affordable housing.

Men too may face unique and particular risks during and after disasters, with their masculinity and socially and culturally constructed roles feeding into vulnerability. In Chicago 1995, following a deadly and prolonged heatwave, the majority of recovered

[106] Delving into this problem, Enarson (2014, p.38) argues:

> Embedded masculinist bias has promoted a "hard" or engineering-related approach to mitigation, top-down notions about knowledge "transfer", and risk-reduction measures derived less from values and interests of women and men in risky environments than those of military business interests, and administrative and political elites. Not incidentally, lack of critical gender analysis has also built a policy environment in which security, disaster, and climate work are ostensibly "gender neutral"... Hard-won disaster experience amply demonstrates the point that women's recovery, hence full family and community recovery, is constrained at foundational levels by embedded male privilege in scientific theory and policy worlds, in the practice and logic of disaster management, and inside the home.

unclaimed bodies belonged to poor African-American males who lived alone, something which Enarson (2014, p.45) attributes partially to the "isolation of gender." Post-disaster/response roles predominantly filled by men also disproportionately expose men to disaster-related occupational hazards—Enarson (2014, p.45) offers the example of the nuclear disaster in Chernobyl where many male responders were exposed to high levels of radiation, and Fukushima, where many men were also exposed to radiation and may yet experience complications as a result.

6.2.2 The Digital Divide and Social Media

The digital divide is a term that essentially refers to gaps in access, proficiency in using, and opportunities for meaningful engagement with modern ICTs, including the internet (Floridi, 2002; Moss, 2002, p. 161). The digital divide is a phenomenon which is defined by unequal adoption of IT services within and between nations[107]—the result of this phenomenon is the exclusion of persons from maximum engagement with the infosphere and all it has to offer (OECD, 2001; Floridi, 2002; Moss, 2002).

The Pew Research Center has conducted research that examines internet access and smartphone ownership across 40 states and territories across three levels of economic development (developing economies, emerging economies, and advanced economies) that offers a useful comparison of internet access and habits between states, and illustrates further the extent of the digital divide on a macro level (Poushter, 2016).

Pew research found that while the proportion is growing, in 2015 54% of persons in the surveyed emerging and developing economies reported using the internet at least occasionally, in contrast to 87% in advanced economies (Poushter, 2016). The quantitative gap is 33%, and is particularly extreme in sub-Saharan African states where 25% of those surveyed were internet users (Poushter, 2016). This research also found that 37% of persons in emerging and developing economies owned internet capable smartphones, and "overwhelming majorities" of people across all nations owned some form of mobile device (Poushter, 2016). Those most likely to own a smartphone were

[107] Between nations, the gulf between internet users and non-users can be quite vast; consider that in Eretria in 2015 1.08 persons per 100 population used the internet compared to 98.20 in Iceland (World Bank, 2017). Even examining economically advanced States using aggregate statistics unmasks discrepancies in engagement with the internet among their populations. In the United States for instance the 2015 figure for number of internet users per 100 population was 74.55; and looking towards the project partner countries of the Slándáil project for particularly relevant examples, the figure for Ireland was 80.12, for the United Kingdom 92, for Italy 65.57, and for Germany 87.59 (World Bank, 2017).

found to be more educated people on higher incomes, and younger persons (aged 18 to 34) are most likely to own a smartphone and use the internet across all states.

This Pew research also found gender to be a substantial determinant in internet usage and smartphone ownership in 20 of the 40 surveyed states, with women lagging behind men by significant margins, particularly in developing and emerging economies (Poushter, 2016).

As an important concern of this research is the engagement of internet users with social media, it is also important to review the demographics of social media users. Pew has also conducted extensive research on this (Poushter, 2016). Pew found that 76% of internet users across all surveyed countries were social media users, and those who were most prolific were from states with lower access rates, that is, internet users in developing and emerging economies were found to be more likely to use social media than their counterparts in advanced economies (Poushter, 2016).

While Pew did not conduct extensive research on the particular user demographics and memberships of social media users internationally, it did conduct extensive research on American users (Perrin, 2015; Greenwood, Perrin and Duggan, 2016). Pew found that in 2016, 79% of all adult internet users use Facebook (and 68% of all Americans), in contrast to 24% online adults for Twitter (and 21% of all Americans)(Greenwood, Perrin and Duggan, 2016). Membership of these sites broadly favours younger cohorts without striking differences between other cohorts (Greenwood, Perrin and Duggan, 2016). Other popular sites include Instagram (32% of those online), Pinterest (31%), and Linkedin (29%) (Greenwood, Perrin and Duggan, 2016).

In the US, social media usage is not marked by sharp unequal usage by broad ethnic and racial characteristics, though Black Americans fall noticeably behind in social media membership: in 2015 65% of White Americans used social media, 56% of Black Americans, and 65% of Hispanic Americans (Perrin, 2015).

The reasons for the digital divide are varied and complex but broadly reflect the reasons for the particular plight of persons bearing the personal characteristics referred to in the preceding subsection—social, economic, political and cultural forces have a large impact on who uses the internet.

Telecommunications policy is one factor which influences the digital divide; liberalisation of telecommunications services and the resultant absence of monopolies

has been argued to result in greater access due to lower prices and improved service quality (OECD, 2001; Guillen and Suarez, 2006, pp. 685–686). The high cost of internet access can be prohibitive for those on low income (OECD, 2001).

Education can be a significant determinant of internet access too, with literacy being a basic requirement for engaging with internet services and with higher levels of education contributing to income earning potential and therefore the ability to afford access to internet services (OECD, 2001; Warf and Vincent, 2007).

The presence or absence of democracy is another important factor that influences the digital divide; where governments are authoritarian or totalitarian and seek to regulate or control their citizens' access to information, and what kind of information they can access, engagement with the internet's full range of services will be hampered (Guillen and Suarez, 2006, pp. 686–687; Warf and Vincent, 2007).

Geography matters too; urban areas have greater access to and better quality internet services than rural areas (OECD, 2001).

This combination of factors will hold a particular weight on influencing women's access to the internet. The global digital divide between men and women is estimated to be 200 million in favour of men, a gap which is fed by the poorer circumstances of women relative to men in terms of income and opportunities—prohibitive access costs to the internet exclude many women whose situations are already marginal, and cultural and political factors also contribute to a lack of representation of women online where discrimination is acutely embedded in the social and political structures of some societies (the total proportion of Arab women online for instance is thought to be 20%) (Warf and Vincent, 2007, pp. 88–89; Broadband Commission Working Group on Broadband and Gender, 2013).

The cumulative effect of the digital divide is the disempowerment of those without access to the internet and modern ICT services. They are robbed of opportunities for expression, association, education, and for commerce. To some extent being on the wrong side of the digital divide is self-reinforcing; by being unable to engage with the internet and draw from its fruits, the digitally excluded are limited in their ability to build their capacity to engage with the internet—by being excluded from a potential source of education and income they are excluded from avenues that can assist them in accessing

and meaningfully engaging with costly ICT services. To that end, the problem of the digital divide is comparable to an ouroboros, a serpent eating its own tail.[108]

6.2.3 Consideration of Inequality in Emergency Response

It is evident that natural disasters are not necessarily equal opportunity events; the weight of their impacts can and does discriminate based on pre-existing vulnerabilities to external shocks that further marginalise those who are already living on the margins. Neither are those living on the margins doing so by pure chance; the social, political, cultural and economic configuration of societies are major determinants of peoples' livelihoods and outcomes based on personal characteristics—they create and reinforce divides. Those who are in disadvantaged circumstances in their everyday existence will bear a disproportionate burden in the event of a natural disaster, they may be more likely to be directly impacted, impacted more heavily, and will struggle more harshly to recover after the event. Those whose identities are composed of the intersections of personal characteristics that render persons more vulnerable to disaster are probably even more likely to struggle to survive and recover from disasters with dignity.

All of this has implications for natural disaster planning and response, however, the long-term response by society and the political institutions through which society acts should be to strengthen the capacity of the marginalised—to build their resilience to external shocks through broad developmental interventions.[109]

The more urgent concern in the event of a natural disaster is how resources should be allocated upon the immediate impact. Resources are not in infinite supply and must be allocated by need. A persuasive case can be made that the traditionally marginalised, based on their exposure and vulnerability, should be the prioritised beneficiaries of emergency response. In what follows a case will be made, using IE and Fiduciary Theory, that marginalised groups should be prioritised, within reason, in emergency response.

The stratifications that compound vulnerabilities to disaster are also broadly reflected in the digitally excluded. The less educated, those on low income, women (where political

[108] Floridi (2002, p. 3) offers an insightful remark on this digital divide: "[t]he DD disempowers, discriminates, and generates dependency. It can engender new forms of colonialism and apartheid that must be prevented, opposed, and ultimately eradicated."

[109] A complex variety of interventions would be required, including a focus on education—on a general and disaster-preparedness related level—social welfare, employment opportunities, initiatives to combat gender discrimination at all levels of society, adequate housing, infrastructural protections against hazards, the list goes on.

and social climate is particularly restrictive), and intuitively those with disabilities, have less access to the internet. They cannot broadcast their voice across the infosphere. Consequently, the digitally excluded will be invisible to any technology that is tasked with finding the signals in social media amongst the noise—signals will less likely be composed of the cries for help and reports of the marginalised. It stands to reason then, that if inadequate precautions are taken, a system that is incapable of adequately reflecting the needs of the marginalised or delivering those needs as messages to emergency responders, stands to bias the pool of information in favour of the more privileged, and therefore inform the allocation of resources in disaster response in a manner that excludes those who are already marginalised.

6.3 Justice in the Infosphere

Having established the prevailing context of inequality in natural disasters and having outlined the concept of the digital divide and some of its causes and consequences, it is time to explore the theory, using Information Ethics as a foundation, that can assist in a normative analysis of inequality in response to natural hazards that uses data obtained from social media, and uncover the moral implications of systems that are implemented in such a scenario.

6.3.1 Capability and Justice

As broadly discussed in Chapter 2, Information Ethics promotes a principle of ontological equality, in-keeping with its holistic environmental approach to ethics it promotes the idea that essentially all information objects hold a minimal moral worth, and as such are worth some minimal respect. This is not to say that the importance of humanity is displaced in ethics, for not all things are alike in their dignity, and those beings with the agency and intention to influence and shape the infosphere around them possess the most dignity (Floridi, 2013).[110]

[110] As Floridi (2013, p. 76) argues in support of this:

> Intuitively, from the point of view of the infosphere and its potential flourishing and enrichment, responsible agents, such as human-beings, full-AI agents, extraterrestrial minds, angels, and God, have greater dignity and are the most valuable informational entities, deserving the highest degree of respect, because they are the only ones capable of both knowing the infosphere and taking care of it according to the conscious implementation of their self-determined projects by increasing or decreasing the anti-entropy levels of their actions.

The human ability or latent capacity to consciously implement actions in the infosphere that can appropriately regulate entropy is what gives human dignity its prized status and high moral value—it is because humans can consciously alter and shape existence in ethical ways that grants them special status above all other informational entities (Floridi, 2013).

A morally responsible agent's capabilities must be nurtured and supported so that they can contribute to the flourishing of the infosphere. The morally responsible agent will need the autonomy and power to implement its actions. Before the infosphere as a whole can flourish, the morally responsible agent must be able to flourish. The relationship between the infosphere as a whole, and its responsible constituents, is symbiotic and both require each-other for nourishment.

If dignity is a function of responsibility and (moral) agency, and the flourishing of the infosphere can be contributed to from this agency, then there is a broad moral imperative to support this agency through:

- Supporting agent *autonomy*—a responsible agent must have *reasonable* freedom and choice.
- Supporting agent *interactivity*—a responsible agent must be able to meaningfully interact with the infosphere.
- Supporting agent *adaptability*—a responsible agent must be able to change and grow in response to its environment.
- Supporting the *Good Will*, or preventing intentional evil—a responsible agent should have their ethical knowledge cultivated by appropriate education and should be instilled with the desire to be caring, or to be a beneficent agent.

It might be noted that humans possess particular dignity not because they implement actions, but because they *can* implement them.[111] There is a particular dignity to

[111] Of course there will always be particular inescapable impediments to responsible agency, from infirmity to various disabilities both physical and mental. All conscious, thinking human beings have varying capacities for autonomy, interactivity and adaptability that should be nurtured. The development and growth of all conscious or potentially conscious humans should be supported regardless of the maximum levels of autonomy, interactivity, and adaptability that they can achieve, as they will all have varying capacities to help the infosphere flourish, and so too must the infosphere be adapted to help them flourish. It may be that the acute physical or mental limitations of some prevent them from exhibiting or being capable of any substantial responsible agency. Such extreme positions limit the dignity that they hold, but entitles them to no less support in developing and actualizating their capabilities, as such capability enhancement contributes to their flourishing, an important principle of information ethics—on the contrary, and as will be argued later, their extreme positions demand that they receive more support than

humans as a general class based on their capacity or potential to contribute to the infosphere. Any obstacles or impediments to the responsible agency of human beings that prevent them from implementing actions that can help both themselves and the wider infosphere from flourishing should be considered possible sources of entropy.[112] Unjustified impediments to responsible agency are impediments to human dignity, ergo they should be removed from the infosphere.

Due to its sometimes Aristotelian (though more ecumenical in its inclusion of all Being and not merely the individual) approach to flourishing and generally conforming as an ethics of flourishing, IE is "sympathetic" to Capability Theory, a theory advanced and developed notably by Amartya Sen and Martha Nussbaum (Sen, 2001; Nussbaum, 2003, 2008; Bynum, 2006; Johnstone, 2007, p. 99). Examining Capability Theory from here will prove instructive in determining a fuller account of equality and justice that is compatible with the ontology and principles of Information Ethics.

In his work on capability theory, Sen (2001, p. 14) emphasises the importance of freedom[113] and "free agency", arguing that:

> Expanding the freedoms that we have reason to value only makes our lives richer and more unfettered, but also allows us to be fuller social persons, exercising our own volitions and interacting with—and influencing—the world in which we live.

For Sen (2001, p. 17), freedom involves both the autonomy to implement actions and decisions as well as the opportunities that people have in life. Obstacles to freedom are termed "unfreedoms", which are the result of inadequate processes[114] and inadequate opportunities for the achievement of goals (Sen, 2001, p. 17).[115]

Persons with adequate freedom are able to achieve what Sen (2001, p. 75) calls "functionings", that is, things that people: "...may value doing or being" that can vary

---

those in a greater position to realize their capabilities. Extreme examples arise where humans have virtually no capacity for autonomy, interactivity, or adaptability and much less responsibility (e.g. the brain dead), however as Floridi (2013, pp. 114-122) extensively argues, such persons are still worthy of respect with regard to the principal of ontological equality.

[112] Consider hunger or poverty, or institutionalised discrimination that prohibits persons from exercising their freedom based on personal characteristics.

[113] Sen (2001, p. 10) lists categories of freedoms as political freedoms, economic facilities, social opportunities, transparency guarantees and protective security, each of which advances individual's capabilities.

[114] Sen (2001, p. 17) offers the example of violation of voting privileges.

[115] Sen (2001, p.17) offers the example of the absence of the capability to escape premature mortality.

"...from the elementary ... [functionings], such as being adequately nourished and being free from avoidable disease to very complex activities or personal states, such as being able to take part in the life of community and having self-respect." Where a person can plausibly achieve between different functionings, they have capabilities, which Sen (2001, p. 75) argues are "...the substantive freedom to achieve alternative functioning combinations." Those with capability have the option of realising different functionings.[116]

Simply put, functionings are the valued things that we do or can be, whilst capabilities are options for functionings that we have at our disposal. The marginalised will have less functionings and capabilities than the more privileged.

Sen (2001, p. 284) argues that those who do not have capabilities (therefore meaningful choice or substantive freedom), such as bonded labourers and girls in repressive societies, cannot be responsible agents—for responsibility requires freedom.

Martha Nussbaum (2003) finds numerous faults in Sen's work and arguably improves upon it by defining its limits and supplementing it with more content. Nussbaum (2003, p. 46) finds it problematic that Sen believes that freedom is itself always good, even though it may be badly used, as "...so much depends on how one specifies the freedoms in question."[117] A fetishistic and uncritical belief in freedom as an unqualified good is dangerous in itself, and has been to some extent the reason for feminist critique of the ideals of political liberalism, which some argue is itself dangerous for failing to intervene

[116] Illustrating this, Sen provides the following example:

> ...an affluent person who fasts may have the same functioning achievement in terms of eating or nourishment as a destitute person who is forced to starve, but the first person does have a different "capability set" than the second (the first can choose to eat well and be well nourished in a way the second cannot).

[117] Nussbaum (2003, p. 46) argues convincingly that:

> Some freedoms include injustice in their very definition. Thus, the freedom to rape one's wife without penalty, the freedom to hang out a sign saying "No Blacks here," the freedom of an employer to discriminate on grounds of race or sex or religion—those are freedoms all right, and some people zealously defend them. But it seems absurd to say that they are good per se, and bad only in use. Any society that allows people these freedoms has allowed fundamental injustice, involving the subordination of a vulnerable group. Of other freedoms, for example, the freedom of the motorcycle rider to ride without a helmet, we should not say, "good in itself, bad only in use," we should say "neutral and trivial in itself, probably bad in use."

in violent actions against women for instance, which have gone unchallenged due to the public and private divides that liberalism promotes (Adam, 2005).

Nussbaum's (2003, p. 49) opinion on the good of freedom is highly qualified, and she believes that promoting freedom as essential to the good life insufficiently respects the values of different societies that see value in living under authoritarian religion, for example, and therefore offers insufficient respect to pluralism—she argues that "[w]e should respect people who prefer a life within an authoritarian religion (or personal relationship), so long as certain basic opportunities and exit options are firmly guaranteed." This is a reasonable view, it minimises the essentialism of freedom (which can be good or bad) while acknowledging that it remains important, and that persons should have at least an opportunity to enter in and out of environments where their freedom will be constrained.

In addition, Nussbaum (2003, p. 44) criticises Sen's failure to provide a list of capabilities in his own attempt to allow plural outcomes and rankings of capabilities, which he fears would "inhibit democracy." So too does Sen fail to adequately acknowledge that some freedoms limit others; Nussbaum (2003, p. 44) gives the example that "[t]he freedom of rich people to make donations to political campaigns limits the equal worth of the right to vote."

Nussbaum gives content and form to capability by endorsing a list of ten central human capabilities, or political entitlements (Nussbaum, 2003, 2008). Nussbaum's list is succinct and minimal which leaves it open to some interpretation to a given society, in order to respect pluralism, however she (2003, p. 48) forcefully argues that these ten capabilities are essential across time and space:

> ...the value of respect for pluralism itself requires a commitment to some cross-cultural principles as fundamental entitlements. Real respect for pluralism means strong and unwavering protection for religious freedom, for the freedom of association, for the freedom of speech. If we say that we are for pluralism, and yet refuse to commit ourselves to the nonnegiotability of these items as fundamental building blocks of a just political order, we show that we are really half-hearted about pluralism.

The list of central human capabilities endorsed by Nussbaum (2003, p. 41, 2008) includes Life; Body; Bodily Integrity; Senses, Imagination, and Thought; Emotions; Practical Reason; Affiliation; Other Species; Play; and Control Over One' Environment.[118]

This list is useful, and it stands to reason that for a responsible agent to be able to fruitfully contribute to the infosphere, they will require the achievable functionings that unlock these ten capabilities—having these capabilities means that a responsible agent should have the opportunity to flourish, and in conjunction with other agents, particularly within a multi-agent system, and help the wider infosphere to flourish too. Capabilities, it should be noted, are supported both by individual states of action and being (that is, the functionings of the individual), and the environment that they inhabit, which will either facilitate or prevent functionings and could shrink capability as a result (an autocratic regime for instance might censor free speech, which would be a direct threat to Senses, Imagination, and Thought)(Johnstone, 2007; Nussbaum, 2008). Again, the relationship between individual and environment, or agent and infosphere, is symbiotic and the flourishing of both mutually reinforcing.

This capabilities approach removes the emphasis that literature often places on social goods, focusing instead on human capacities to achieve certain ends. It to some extent shifts the focus from what people have, to what they can do or can do and be (Sen,

[118] To elaborate more on Nussbaum's (2003, p. 41, 2008) list of capabilities, they are:

1. Life: being able to live a life of normal span
2. Bodily Health: being able to exist in good health.
3. Bodily Integrity: being free from all forms of physical violence, and having options for sexual satisfaction.
4. Senses, Imagination, and Thought: essentially being able to use the mind fruitfully, in a manner informed by far ranging education, and being able to act on the thoughts of the mind such as through political and artistic speech as well as religious exercise. Being able to experience pleasure and avoid pain.
5. Emotions: being able to form emotional bonds and relationships with others and being able to experience the spectrum of emotions, and not being burdened by fear and anxiety.
6. Practical Reason: "[b]eing able to form a conception of the good and to engage in critical reflection about the planning of one's life."
7. Affiliation: being able to live and work socially, have regard and compassion for fellow human beings and interest in their well-being. Being treated as an equal by others and holding self-respect.
8. Other Species: being able to live in the biosphere with concern and respect for its constituents.
9. Play: being able to enjoy recreational activities.
10. Control Over One's Environment: being able to have involvement in political processes and being able to have having access, ownership, and security over material goods.

2001; Nussbaum, 2003). Using Capability Theory can help provide more substance and clarity to the IE approach taken; true injustice occurs where persons lack in capability whilst others do not—low capabilities limit responsible agency and the flourishing of the responsible agent, whose ability to act upon the world in a meaningful, or moral, way is also limited.

When persons in a society have unequal capabilities, some will be worse off than others, as was extensively outlined in the preceding section. Entropy will disproportionately affect those with lesser capabilities, with less functionings—consider those without transportation functionings, who may have no choice but to remain behind in a disaster whilst those who do have that functioning can decide whether or not to leave before or shortly after the disaster strikes; the person without transport functionings may find themselves without the overall capabilities to safeguard their life, health, or bodily integrity. That the structure of society exposes some moreso than others to this entropy is an injustice, and this entropy further reduces their capabilities, their general flourishing and their capacity to contribute to the flourishing of the wider infosphere, is an injustice. In a just society, "...the opportunity to develop and express capabilities is provided to all" (Johnstone, 2007, p. 77).

It follows that in the process of disaster and emergency planning, the capabilities of all persons should be analysed, and disaster planning, mitigation, and response strategies be tailored to empower those with less capabilities and therefore greater vulnerability to disaster.

The digital divide is both a cause and consequence of a lack of capabilities—persons can lack the functionings required to engage with ICTs (they can lack the resources, skills and access) and without access to ICTs, they lack the ability to expand their functionings and therefore capabilities (without sufficient access to ICTs, they lack the ability to build their technological literacy). That digital aspect of the infosphere, cyberspace, is a unique territory in itself where capabilities can be nourished, where people can exercise political and artistic functionings, socialise, fall in love, play, and experience emotion. To be shut out of this realm of the infosphere shrinks capability, and limits important freedoms.[119]

[119] With regard to the pluralism promoted by the capability approach, and generally endorsed by information ethics, digital inclusion and the functionings and capabilities it promotes may not necessarily be considered functionings of value in a given society; some people may not perceive

The question now arises as to whose responsibility it is to ensure basic capabilities. Multi-agent systems are those which are best placed to address inequalities of opportunity that reduce capabilities—whether these are states, or multi-national entities such as the United Nations. It is the responsibility of such agents, with their combined knowledge and power, to reshape the environment into one which can maximise functionings of the individual agents of which they are composed, to allocate resources based on need (distributive justice) and empower individuals (Johnstone, 2007). Those with the least capabilities should be prioritised patients of corrective action.

Of course this is not to say that smaller entities do not hold responsibility for exclusions for which their creations might be responsible, where their creation may reduce the capabilities of others. As an example, software developers may create a software product that's use becomes itself an important functioning, and lack of access to which may reduce the capabilities of excluded users.[120]

In summary, human beings have a special status of dignity based on their responsible agency which allows them to make meaningful contributions to the infosphere, to help it flourish. Obstacles to this agency may be sources of entropy (that should be removed). All humans should be given minimum opportunities to exercise agency, and this will require the removal of barriers to their agency and its enhancement. Agency can be enhanced through functionings, and a flourishing, responsible agent will be able to achieve functionings necessary to have capabilities as listed by Nussbaum.

---

value in digitally enhanced capability. A hermitically sealed off tribe living prehistorically in the deepest regions of the Amazon, for example, may not consider the functionings associated with ICT use essential for their well-being, they may not value such beings and doings and as such the absence of such functionings would not in their knowledge constrain their capabilities. Nevertheless, as the resources and technology of the digital space represent such an immense opportunity for development, persons should have at least the opportunity to engage with them—a person arguably has fuller capability if they have the option of opting in or out of the digital space, rather than not having the option at all.

[120] Individuals as responsible agents with power over other individuals too can be responsible for the capabilities of others; consider the relationship between parents and children; it could be argued that particularly fortunate agents with much power and resources should also make contributions to the capability development of others, such as through charities. It might be too that an agenda of appropriate justice by the MAS can only be set by the lobbying efforts of individuals and groups—therefore, while the state may be the best placed to enhance capabilities, it is up to the constituting agents of the state to apply pressure for it to pursue just actions (should it be failing to do this without external pressure). Again distributed morality is important, aggregate actions can combine at state MAS level and achieve moral outcomes in the form of policies and actions that achieve the requirements of justice.

Finally, the world is not Utopian and resources are not unlimited. Though when speaking of capability one shifts discussion from resources to empowerment, resources remain important, especially in the context of natural disaster when they are stretched and diminished. The following section will appeal to the logic of David Parfit's (1997) Prioritarianism to justify the diversion of resources to the least capable and supplement the discussion already made here.

6.3.2 Priority, Capability, and Emergency

The argument offered here holds that inequality (of outcomes, at least) is not evil in itself, it is not an untenable telic egalitarian[121] position that states that all humans should have perfectly equal circumstances, however it does hold that all humans have a basic moral equality and that their responsible agency should be supported to the extent that they have central capabilities necessary for living a life of responsible agency (within the context of their society and its values), and that conditions obstructing the fulfilment of these capabilities are essentially sources of entropy. The continued advancement of capability is morally good; the continued flourishing of human beings and the infosphere generally must be an ongoing project, however there is no requirement that all people be exactly equal in all circumstances, merely that people have the opportunity to grow and flourish with the assurance that this growth and flourishing will be supported within the minimal requirements of having a worthwhile life.[122]

As demonstrated in the present chapter, unequal functionings and capabilities based along the arbitrary lines of personal characteristics tend to make natural disasters discriminatory events. Unequal capabilities render those with less functionings

[121] Of Telic Egalitarianism, Parfit (1997, p. 204) argues: "[s]ome egalitarians believe that, if people were equally well off, that would be a better state of affairs. If we hold this view, we can be called *Teliological*—or, for short, *Telic*—Egalitarians. We accept The Principle of Equality: It is in itself bad if some people are worse off than others." The position can be used to make extreme points, such that in a society where half of the population were born without eyes, those with two should be forced to donate one, and generally can follow the kinds of logic as explored in Kurt Vonnegut's (1994) short story, Harrison Bergeron, which depicts a dystopian society wherein all those with natural talents are encumbered or otherwise disadvantaged, or levelled down, so that all may be equal (Parfit, 1997). While such logic flows from the principle, admittedly few (if any) reasoned philosophers have advocated it (Gosepath, 2011). The *reductio ad absurdum* stands however as highlighting the weakness of Telic Egalitarianism.

[122] This is not to say inequality of circumstances is not an outrage, that it is not outrageous that men dominate senior corporate and political worlds as women struggle to ascend corporate ladders or exert political influence, or even struggle supporting a basic existence as the case may be. The evil is not that outcomes differ, the true evil is in the barriers to capability that reduce different peoples' opportunities for flourishing, that reduce the possibility that people have an equal chance at obtaining the same outcomes.

disproportionately vulnerable to the deleterious effects of natural disaster. It is an injustice that societies are arranged that this is so, however the evil is not the outcome, but the arrangements that lead to such outcomes. The lack of capabilities of some in disasters combined with the lack of resources to respond to disasters by responding agencies presents a dilemma for these responding agencies—to whom do we allocate resources in emergency, and why? A utilitarian logic in such circumstances may be tempting (Zack, 2010). Distribution based on maximum utility can be persuasive, but perhaps also callous and unfair. Depending on how utility is measured (perhaps as an individual's capacity to contribute to the economy), decisions can be made that favour those who already have high capabilities at the further expense of those without the same capabilities. Perhaps resources can be allocated where they are thought to save the most lives (which is perhaps fairer), and are distributed amongst areas within the epicentre or initial impact zone of a disaster. This still leaves something to be desired if the case is that such areas happen to be dominated by highly capable populations, it may be that many have already evacuated or their homes have proven to be resilient to impact—this would be an information intensive approach requiring up to the minute data in pressurised events that require quick action; it leaves open the possibility that people's personal characteristics will be weighed against them.[123] The possibility that utilitarian reasoning can be useful in such circumstances will not emphatically be rejected, however a more humane form of reasoning compatible with the Capability Theory modified Information Ethics approach can be found in Prioritarianism, and will thus be supported here.

In the scheme of resource allocation, the priority view holds that "...benefiting people matters more the worse off these people are" and distribution should be according to need (Parfit, 1997, pp. 213, 216). For the purposes of this research, those who are worse off or most in need should be considered as those with less functionings or low capabilities, and therefore less opportunity for expressions of responsible agency. Extending particular assistance to those with the lowest capabilities, is a matter of "...*humanitarian concern*, a desire to alleviate suffering" (Gosepath, 2011). The urgency with which someone deserves to be helped is based on how poor their circumstances, the poorer their situation (in terms of capability here), the more deserving they are of

[123] Resources would less likely to be distributed to areas known to be populated by the aged, or those with ill-health, as their survival chances with or without intervention may be viewed less favourably.

help even if "...they can be less helped than others in the process" (Parfit, 1997; Gosepath, 2011).

Using the logic of Prioritarianism, in a disaster response situation, let us say a massive flood, emergency managers should allocate resources to those with pre-existing low capabilities (those resources may be rescue vehicles, food and medical supplies) even if the case might be that those same resources, directed at individuals with more capabilities would be of greater utility.[124] Then assistance should be directed at those who broadly need it more, even if others would benefit from it more, unless compelling reasons can be offered for prioritising those less in need (Parfit, 1997).[125]

There is little reason that prioritisation cannot be granular or context specific, for instance, particular resources can be distributed to those without the particular functions and for whom receipt of these resources would grant central capabilities. Take for example transportation as a functioning—emergency managers could prioritise dispatching rescue vehicles to those who do not have transportation leading up to a disaster,[126] or in terms of personal safety and shelter in more extreme circumstances, could prioritise women who may lack these functionings by allocating them gender segregated shelter and additional security personnel.

Such decisions can be made with the assistance of Census statistics or other research data, where they exist, which can offer significant insight into spatial concentrations of various population groups that can essentially include functionings (whether or not a car is owned) and personal characteristics (class, gender, age, and single women with children for example).

Prioritising based on need will not be a simple task, and would require substantial groundwork ahead of an emergency in order to determine the demographic composition, and perhaps types of accommodation which persons occupy in a given location, in advance of an emergency. Essentially, emergency management would

[124] Perhaps the poor and sick are less likely than the young and healthy to respond to these resources, and therefore stand to benefit less—it is the extremity of their situation that gives them a claim to priority.

[125] It is important to note that there can indeed be situations where advantaged groups need it more; presume for instance that the impacts of a disaster are highly concentrated in an affluent area, but surrounding disadvantaged areas are minimally impacted. Answers may not always be clear, though emergency managers ought to consider the impacts on the disadvantaged before committing resources.

[126] This decision might be more difficult where it can be assumed, or is known, that a disaster has destroyed transportation infrastructure.

benefit from thorough needs based assessment during disaster planning based, as closely as possible, on the capabilities of a population.

The logic of Prioritarianism is compatible with and supplements that of IE and Capability Theory, and is useful to rely on in contexts of resource shortages. In Information Ethics, dignity is a product of responsible agency, which is difficult to achieve when capability is low. Those who have low capability then, suffer from jeopardised dignity and low functioning. They are victims of entropy, and are in need of capability enhancement so that they may flourish. Where resources are low, they should be directed first to those with low capabilities, who require them as a matter of urgency and are acutely vulnerable to external shocks that can further reduce their capabilities. In making this decision, resources will be directed away from persons with high capability, who may in fact benefit from them more. Information Ethics is a patient centred ethics of care however, and demands a particular compassion for those who are most vulnerable. By diverting resources from those termed here as being high capability, it may be so that emergency managers are allowing them to succumb to entropy. Evil in the infosphere can not be entirely prevented, and difficult decisions have to be made. Saving a life is good, and it is probably best that those least likely to be able to save themselves, or recover sufficiently from a disaster, be the main beneficiaries of emergency response—this is a good act, as well as an act of care and compassion.

6.3.3 Slándáil-type Systems and Justice

At this stage, it is appropriate to investigate the implications of Slándáil, and by extension similar systems, for human capability. The ideal outcomes assisted by Slándáil can promote human functionings that build capability. As outlined in Chapter 4, the transmission of timely information to emergency managers, in an ideal scenario, will result in a rapid dispatch of resources to those in distress, potentially resulting in saved lives. In such a case, functionings such as personal safety, or transportation, might be necessary for Life capability at least, and the delivery of resources can provide these functionings. Systems like Slándáil then, can relay to emergency managers information about credible threats to capability, and enable emergency managers to respond to these threats. The Slándáil system and wider social media, also enable other sets of functionings that can contribute towards capability. By tweeting (for example) towards a goal, particularly in the context of disseminating disaster information or reports, and contributing to the MAS of agents involved in disaster response, humans are

demonstrating efficacy at other capabilities, such as Senses, Imagination, and Thought (creating messages about disasters); Affiliation (connecting with others and the disaster affected); and Other Species (concern for the disaster affected). The form of DM explored extensively in Chapter 4 demonstrates the potential for capability enhancement.

Of course, on that note, functionings are required to contribute to the DM of the Slándáil empowered MAS, as also extensively discussed in Chapter 4. To contribute, one must have a range of functionings, from as basic as literacy, to IT literacy (a social media account and ability to use it) as well as the supporting conditions (such as an internet connection, and devices that can access the internet), themselves requiring functionings (states that generate income such as employment). Persons without these functionings will necessarily be at a disadvantage, and may not be in a position to contribute directly to the pool of social media information available to Slándáil, similar systems, and by extension emergency managers. As the Slándáil system has the power to contribute to the protection of capability (saving lives), this lack of functionings can result in lost capability. Returning to the example of Alice and her flooded town can illustrate the problems that arise from this dilemma.

Suppose again that Alice's town has flooded and she is trapped in her home. The economic environment is unfavourable, and Alice is reliant on modest social transfers from the government. Alice is a single mother with a young son, whose well-being is a major priority for Alice. Alice spends much of her income on rent and utilities, and on feeding and clothing her son as well as paying for his school books. Alice does not have the budget for what might be considered luxury goods, her mobile phone is basic and is decidedly not a smartphone, neither does she have a computer or an internet connection. She is called to the balcony by her son who points out the torrents of flood water crashing against her neighbours apartment complex, which is old and not built to a great standard. On the first floor of this complex, she sees her neighbours at their balconies looking on fearfully. She immediately moves to her landline phone and attempts to dial 911. The lines are engaged. Alice has few options. She has limited functionings, as do her neighbours. The emergency response agencies are responding to situations elsewhere and she has no way of alerting them to this rather serious threat. Eventually her neighbours' apartment complex collapses, killing most inside. The tragedy

of the Good Will has occurred, in part, because Alice (and her neighbours), have limited capabilities due to their impoverished circumstances.

Variables in this example can be swapped without altering the outcome. Suppose Alice does have an internet connected device (an old laptop) and social media account, but does not speak the common tongue of the state, but a language unrecognised by the Slándáil system, she is an undocumented migrant from a neighbouring country. She might tweet about her situation, but the message could easily be lost and certainly missed by the system's filtering and analysis without another Twitter user helpfully intervening with a translation. We can suppose that Alice has limited literacy and no need of a social media account—she enjoys using the internet to watch videos and such, but does not have the ability to coherently articulate her thoughts in writing. Still she would be unable to broadcast her message to the infosphere.

In contrast, we can suppose that south of the river bank, where the more prosperous live, many are tweeting about their situation, and emergency managers are receiving a steady stream of information from these areas, leaving them with little choice but to allocate resources there whilst they remain in the dark about circumstances north of the riverbank. They can only act upon what they know.

With that said, it may only take one person with access to the internet and an appropriate device to make a report that could be seen by emergency managers. The plight of Alice and her neighbours would be greatly mitigated if even only one made a tweet, especially with an image. A further compounding problem however would be the likely clustering of persons with less functionings into particular geographical spaces, potentially resulting in less in information being available related to the situation in one area when compared to another.[127] Such information grants knowledge, and knowledge with power compels action—if an emergency manager has detailed situational awareness pertaining to an affluent neighbourhood, but little about a disadvantaged area, response may militate in favour of the area where there is a greater understanding of immediate local needs.

These examples illustrate perhaps worst case scenarios, however they do amply demonstrate the potential for injustice to arise from the implementation of systems

[127] The present researcher was involved in a project, Cork City Profile 2014, a geo-spatial socio-economic profile of Cork city, which illustrates concentrations of persons suffering different disadvantages in particular areas (Kelly *et al.*, 2014).

such as Slándáil, their capacity to contribute to biased judgement in disaster response as a consequence of the digital divide. In a worst case scenario, those with low functionings might be ignored while they are the very ones who should be prioritised in emergency response to natural disaster.

Indeed, whilst plausible scenarios have been demonstrated, the reality may be more forgiving. Emergency managers are in receipt of expert information (or "intelligence", which is the preferred term) from a plurality of sources, from meteorological outlets to satellite image services such as Copernicus and as the situation currently exists, will likely trust such information outlets before social media reports. With things as they currently are balanced situational awareness is likely to prevail. Interviewed emergency managers denied that systems such as Slándáil would displace any current information collection processes.

It however remains apparent that IT related functionings will become less and less optional and will became basic requirements for capabilities as time goes by. It is proving necessary to be able to engage with the social media aspect of cyberspace to realise a range of capabilities. Those without such functionings are also likely to lack a variety of others. Section 6.2 of the current chapter demonstrated an overlap between those who fare poorly in disaster, and those who are digitally excluded. The Prioritarian view supported in this research requires that such persons, on the basis of their low capability, are prioritised in natural disaster response, and such prioritisation will be hampered where information relating to their needs is inadequate or unclear. Any technology which stands to make such persons even more vulnerable poses a problem, and should be designed in a manner that mitigates this problem.

There is an incumbent need on software designers to ensure that their creations, which have the potential to compound the consequences of digital exclusion, consider how their creation interacts with the functionings and capabilities of those whose lives their creation will impact. These persons, or larger entities, too are responsible agents with correlative obligations in the infosphere, and should ensure that their creations make the greatest possible positive contribution without causing entropy.

Technical solutions are possible that can mitigate the consequences of the lack of representation of the digitally excluded online. In many countries, pre-existing Census and other data exist to be exploited by systems such as Slándáil, which can be loaded

into GIS based system architectures, such as that described in SIGE. Whilst it is impossible to find the voices of the vulnerable on social media if they are not present there, datasets can be selectable and overlaid on interactive maps that can indicate where society's most vulnerable are. Additionally, variables in such datasets can be combined into new datasets that can convey levels of deprivation. A good example of this is the All Ireland Haase-Pratschke Deprivation Index, an index that allocates a deprivation score to different administrative units based on several variables drawn from Census data (Haase, Pratschke and Gleeson, 2014). Figure 19 illustrates an implementation of this Index on a map of Ireland.

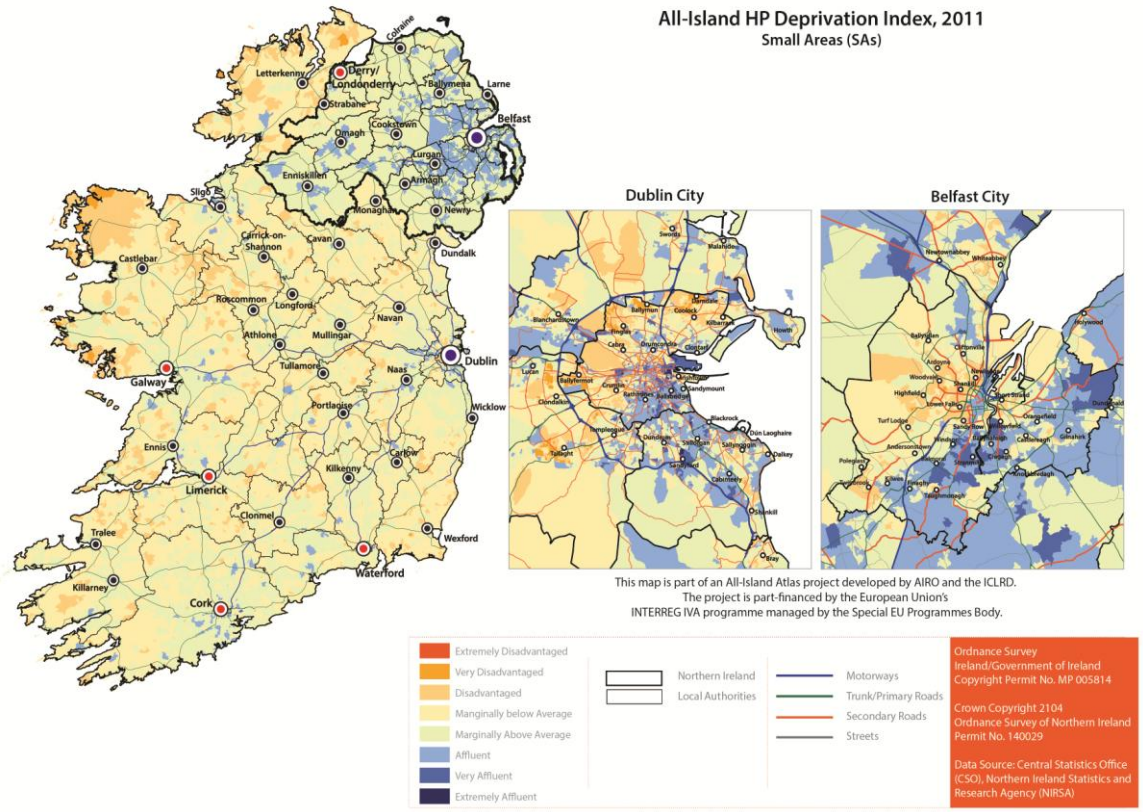

Figure 19: All Ireland Deprivation Index Small Areas (Source: Haase, Pratschke, Gleeson, 2014)

There is also the example of the Social Media Index, which uses the Bonferroni model, as described in Chapter 4, that offers a risk score in a given area based on several variables including social media. Such computational methods can, and should, be adjusted to include variables of social disadvantage, in order to inform emergency managers of locations which are likely to be heavily vulnerable based both on their vulnerability to physical shocks and the socio-economic vulnerability of their

populations. This way, emergency managers' decision making ability will be enhanced in a manner that helps them to make Prioritarian lead decisions that broadly benefit those with the least capability.

Beyond this, it remains important for emergency managers to be cognisant of the particular vulnerability of low capability populations in advance of disasters. Planning is essential, and priorities should be planned to the best extent possible in advance of a disaster. To that end, emergency managers should conduct thorough needs based assessments that account for personal characteristics in society in order to establish the precise needs of different populations in the immediate aftermath of a disaster.

As a final note, though those with least functionings should be the prioritised beneficiaries of emergency response, there may be occasions where natural disasters intensely impact affluent areas but leave disadvantaged areas relatively untouched. It can be assumed the though the inhabitants of such areas are affluent, and likely have high capabilities prior to the disaster, these capabilities will be immediately challenged and their situation will be much more precarious than their disadvantaged counter-parts in at least that moment in time. In such situations the normally advantaged may be those most in need of various resources, and should indeed be prioritised in allocation of those resources. The moral arithmetic involved in resource allocation in the immediate aftermath of disaster is unlikely to ever be easy, however it is important that emergency managers bear in mind the tentative position of the disadvantaged or marginalised, and recognise that their experiences are likely to be more severe, and as such their needs should be reflected in response (as well as planning, mitigation, and recovery) policy and their locations accessible through EMIS so that they are not forgotten in response efforts.

6.4 Justice and Human Rights: A Focus on Discrimination

Human rights offer entitlements to a broad range of rights, with respect to a broad range of interests (including material interests like food and shelter). Fiduciary Theory does not offer a full theory of justice (Fox-Decent, 2011), merely a minimal account that does indicate who should be preferred, and especially considered, in any given state action with broad consequence for its subjects. An intense assessment of all human entitlements relevant to disaster and with implication for Slándáil-type systems would be excessive and unnecessary, therefore this section will opt to focus specifically on the principle of non-discrimination due to its intrinsic importance in matters of justice,

which deals with conflicting claims and matters of equality—arbitrary discrimination is clearly antithetical to resolving competing claims. Discrimination can go against the principle of formal moral equality, can deny the equal dignity of human beings, and can result in decisions that are unfair and arbitrary—in light of this, and with regard to the prominence of the principle of non-discrimination in IHRL, a broad examination of non-discrimination here is fitting.

6.4.1 Fiduciary Theory, Equality, and Non-Discrimination

The state as fiduciary is constrained in its actions by the principle of formal moral equality and the duty of procedural fairness (Fox-Decent, 2011; Criddle and Fox-Decent, 2012). The principle of formal moral equality "... requires fairness or even-handed treatment of persons subject to state power; human rights must regard individuals as equal co-beneficiaries of fiduciary states" (Criddle and Fox-Decent, 2012, p. 55).

The principle of formal moral equality requires that all persons subject to the state's power are regarded with equal dignity, and that each has the same entitlement to freedom and dignity—all those whose interests are held in the public trust are entitled to a regime of secure and equal freedom under the rule of law as administrated by the state as fiduciary (Fox-Decent, 2011; Criddle and Fox-Decent, 2012).

Fiduciary Theory does not prohibit all forms of discrimination based on personal characteristics, "...fair exercises of public and fiduciary power do not necessarily treat everyone equally, they treat everyone in a way that acknowledges the standing of important interests vulnerable to public power," and "...equality does not mean equal treatment, it means equal concern and respect. Relevant differences between two persons or their circumstances can justify differential treatment on grounds of fairness" (Fox-Decent, 2011, p. 183). As such, whilst all under the fiduciary's power can claim rights, depending on their particular circumstances, the fiduciary's obligation to respect or fulfil these rights may vary by degrees. Consider the right to vote, for instance, which is often limited for non-citizens. The inequality here arises based on an important and *prima-facie* defensible (though contentious) distinction, which is citizenship, and the full enfranchisement of non-citizens until they have been naturalised through the citizenship process, might not be regarded as being an unfair limitation to the non-citizen's rights

(who may well still have voting privileges in the state of which they are a citizen)[128] based on the overall grounds of security (for example).[129]

The principle of formal moral equality therefore does not absolutely preclude all forms of discrimination. What remains essential is that discrimination is not based on arbitrary grounds. The fiduciary state must regard the interests of all the beneficiaries of its power, based on their equal dignity and ability to place the state under obligation. Discrimination must be well justified, and a duty of fairness should militate in favour of those who are especially vulnerable to decisions made by the fiduciary, as required by the principle of solicitude. In practice, for the fiduciary to ensure a regime of secure and equal freedom to all, and with regards to the equal dignity of all co-beneficiaries, and their different interests, as well as regard for how their interests may be differently affected by various state actions, many discriminatory policies should be required to benefit those whose interests are particularly vulnerable, or who may need particular support to benefit from a regime of secure and equal freedom. The following subsection will explore this practice of positive discrimination—designed to help those with a more tentative and vulnerable position in society—in more detail as it has been practiced in international law. Suffice it to say in the interim, that where a person is particularly vulnerable to instrumentalisation or domination, the fiduciary should pay special heed of their circumstances and intervene so as to protect their rights, even if such interventions are not made to all equally.[130]

As outlined in section 6.2 of the present chapter, a variety of personal characteristics render some people more than others more vulnerable, and challenge their dignity and the enjoyment of rights to which all are entitled. The fiduciary that rules with solicitude

[128] The continuing relation between expatriot and their state of citizenship means that this state continues to hold fiduciary duties towards the expatriot beneficiary of their power—the logic of these relational duties was explored at length in Chapter 5.

[129] The purpose of this research is not to determine whether there is justice in such forms of discrimination, and therefore neither side of this argument will be endorsed, however it remains an instructive example of an entrenched and systematic form of discrimination rooted in (debatably) reasonable grounds that mean that it is not, *prima-facie*, an arbitrary form of discrimination. This is not to say that it cannot, or should not, be challenged, as compelling reasons can also be offered in defence of voting rights for non-citizens.

[130] Consider equality legislation that protects in particular the interests of minorities, or legislation that deters hate crimes by imposing more serious sentences on offenders in order to act as a particular deterrent to such crimes. Or consider social welfare payments made to persons with disability who are unable to work, in this case the able-bodied are excluded from such particular schemes, but such discrimination is justifiable on the basis that the disabled are in such a vulnerable position that failure to provide financial assistance would threaten their enjoyment of a regime of secure and equal freedom.

should strongly regard their interests in enacting policies, and should pay particular attention to such groups in ensuring their rights. The positive obligations imposed upon a state to ensure the opportunity for all to enjoy secure and equal freedom demands that those in marginal positions benefit from extraordinary measures that ensure their rights where they are especially threatened (particularly in the case of disaster, social and economic rights), which can be much more vulnerable. The rights of some are less vulnerable to going unfulfilled than others, and the fiduciary must ensure that all hold an equal opportunity for the fulfilment of rights, even if this means that more vulnerable groups benefit from positive and negative actions that others do not. Fiduciary Theory does not provide a full theory of distributive justice, however it does support the Prioritarian logic that those who are most disadvantaged (in this case, with regard to their opportunity to enjoy the full range of human rights), benefit from more support than those who are most advantaged (those whose rights are already fundamentally secure may require less positive action for the continued enjoyment of these rights).

In what follows, the international human rights law relating to non-discrimination will be examined, with further reference to Fiduciary Theory. Such an examination will bear out the particular responsibilities of states with regards to non-discrimination as established in international human rights treaties and the case law of the ECtHR.

6.4.2 International Human Rights Law and Non-Discrimination

The first major international human rights instrument to condemn discrimination was the *UDHR* (1948), Article 7 of which states that:

> All are equal before the law and are entitled without any discrimination to equal protection of the law. All are entitled to equal protection against any discrimination in violation of this Declaration and against any incitement to such discrimination.

Building upon this, both the *ICESCR* (1966) and *ICCPR* (1966) have provisions that enshrine the principle of non-discrimination. Article 2, paragraph 2 of the *ICESCR* (1966) for instance, prohibits discrimination on the grounds of "...race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth or other status."

The *ICCPR* (1966) makes numerous references to discrimination, ensuring that all rights enshrined within are to be enjoyed by all regardless of personal characteristics,

particularly in Article 26.[131] The *ICCPR* (1966) also firmly establishes positive obligations to prevent discrimination, with Article 20, paragraph 2 stating "[a]ny advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence shall be prohibited by law." Importantly, the *ICCPR* (1966) Article 4, paragraph 2, explicitly prohibits discriminatory practices in response to emergencies.

These international human rights instruments represent formal articulation of the principle of formal moral equality, the moral standing of human beings, and the necessity of equal rights for all people regardless of their personal characteristics.[132] As non-discrimination is enshrined in the *UDHR*—among the most widely signed IHRL instruments—this also demonstrates that equality and non-discrimination are subject of the overlapping consensus (Donnelly, 2013). Whilst formal articulation and codification of rights is not necessary for the creation of rights and consequent state obligations, it is clear that the precepts of Fiduciary Theory are well accommodated in practical tools that can be used to hold human rights abusing states to account, and to strengthen the position of the international community in protecting human rights as their secondary guarantors.

Turning once again to the *ECHR* proves instructive in examining the application of equality and non-discrimination in the practice of the ECtHR, which has to some extent operationalised the principles of Fiduciary Theory in its case law judgements. In the *ECHR* (1950) discrimination is prohibited by Article 14, and was subsequently re-affirmed in *Protocol No. 12* (2000) of the *ECHR,* Article 1, paragraph 2.

---

[131] The *ICCPR* (1966), Article 26 states that:

> All persons are equal before the law and are entitled without any discrimination to the equal protection of the law. In this respect, the law shall prohibit any discrimination and guarantee to all persons equal and effective protection against discrimination on any ground such as race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth or other status.

[132] Note that the international community's interest in eliminating inequality and discrimination resulted in further international human rights treaties which were formulated with a particular recognition of the historical (and ongoing) plight of women, minorities, and the disabled, and in recognition that distinct measures would be necessary to secure their equal enjoyment of human rights. Such further important contributions to IHRL include the *International Convention on the Elimination of All Forms of Racial Discrimination* (1965), the *Convention on the Rights of Persons with Disabilities* (2006), the *Convention on the Elimination of All Forms of Discrimination against Women* (1979), and the *Convention on the Prevention and Punishment of the Crime of Genocide* (1948).

Discrimination can in essence be categorised into three groups; direct discrimination, indirect discrimination, and reverse discrimination/affirmative action. The following three subsections will examine the distinctions between these three categories of discrimination.

6.4.2.1 Direct Discrimination

Discrimination has been defined by the Court as "...treating differently, without an objective and reasonable justification, persons in relevantly similar situations" (Harris *et al.*, 2009, p. 579 citing *Zarb Adami v. Malta,* [2006]). Direct discrimination can be regarded as being a measure, action, or policy that by design treats persons in similar situations differently.

The Court typically examines complaints of discrimination within the context of complaints of breaches of other rights and to this extent Article 14 is considered parasitic—the Court will often not examine the Article 14 complaint if it has already determined that the article with which it was taken in conjunction was violated (even though the application can be made independently of whether a breach has occurred of the other rights in question)—the Court has been inconsistent in this regard however (Harris *et al.*, 2009, pp. 578–579).

Typically (but not always), the Court will follow four steps in determining whether a violation of Article 14 has occurred (Harris *et al.*, 2009). It will firstly determine whether the discrimination falls within the ambit of another article, secondly it will establish if there was a difference of treatment between persons, thirdly if they are in a similar situation (that is, a comparator is necessary to prove that there was indeed a difference of treatment), and finally it will determine whether or not the difference of treatment was objective and reasonable (Harris *et al.*, 2009; European Union Agency for Fundamental Rights and Council of Europe, 2010, p. 579).[133]

[133] This procedure was clarified in *Rasmussen v. Denmark* (Harris *et al.*, 2009 citing *Rasmussen v. Denmark,* [1984]). In this case, the applicant wished to contest the paternity of children, for whom he was paying his ex-wife maintenance. He had a limited time in domestic law to contest his paternity, and was consequently unable to do so, while his wife, in comparison, did not face the same limitations. In that regard, the applicant complained of breaches of Article 6 (right to a fair trial) and Article 8 (respect for private and family life). The Court applied its procedure as detailed above, and agreed that the case fell within the ambit of Articles 6 and 8. It found that there was a difference of treatment of persons (presumed to be) in analogous situations. The Court had to decide whether the difference in treatment was based on a legitimate aim or if there was "...reasonable relationship of proportionality between the means employed and the aim sought to be realised" (*Rasmussen v. Denmark,* [1984]). In this case, given that there was no

The addition of Article 1 of *Protocol 12* of the *ECHR* (2000) also escalates the scope of non-discrimination by prohibiting it more generally, including by regulating private interactions between private actors—the combination of this and *Article 14* ensures that states are under positive obligation to enforce conditions of non-discrimination across the public and private sphere (Harris *et al.*, 2009, p. 610; European Union Agency for Fundamental Rights and Council of Europe, 2010, p. 73).

6.4.2.2 Indirect Discrimination

According to Harris *et al.* (2009, p.607), "[i]ndirect discrimination results from a rule or practice that in itself does not involve impermissible discrimination but that disproportionately and adversely affects members of a particular group... There does not necessarily have to be discriminatory intent for indirect discrimination."

Indirect discrimination then may occur when persons in different situations receive the same treatment—the ostensibly neutral action or policy in question treats them as alike despite their differing circumstances and disproportionately adversely affects one

---

standard approach to the matter established in contracting states, the Court used its margin of appreciation and accepted Denmark's assessment of the legitimate aim and proportionality of the discriminating measures (the State's argument is outlined in the footnote below), and therefore found no violation of Article 14.

As the Court stated in *Rasmussen v. Denmark* [1984]:

> The Government pleaded that the limited difference of treatment that existed had an objective and reasonable justification. They relied, inter alia, on the following points: (i) the respective interests of the husband and of the mother in paternity proceedings were different: unlike the husband's interests, the mother's generally coincided with those of the child; and it was natural that, in weighing the interests of the different family members, the Danish legislature should in 1960 have taken the view that the interests of the weaker party, namely the child, should prevail...; (ii) the legislature had also regarded it as necessary to lay down time-limits for the institution of paternity proceedings by a husband because of the risk that he might use them as a threat against the mother, in order to escape maintenance obligations; (iii) in deciding whether the national authorities have acted within the "margin of appreciation" which they enjoy in this area, regard should be had to the economic and social circumstances prevailing at the relevant time in the country concerned and to the background to the legislation in question; (iv) Denmark had undoubtedly amended the 1960 Act when this proved to be warranted by subsequent... but it could not be said that the former Danish legislation on this matter was at the relevant time less progressive than that of the other Contracting Parties to the Convention.

person or group above another (European Union Agency for Fundamental Rights and Council of Europe, 2010, p. 29).[134]

[134] The fundamental principles in assessing cases of indirect discrimination are apparent from reading the case of *D.H. and Others v. The Czech Republic* [2007] (Harris *et al.*, 2009). This case related to education legislation in Czech Republic (Ostrava in particular) that resulted in a disproportionally high number of Roma pupils being assigned to special schools in comparison to non-Roma pupils—50.3 percent of Roma pupils were assigned to special schools in comparison to merely 1.8 percent of non-Roma pupils (according to statistics obtained from a survey conducted in Ostrava); other data sources suggested that as high as 70 percent of special school pupils were Roma. Most of the applicants in the case, parents of the children concerned, contested (outside of the formal appeals procedure) the administrative decisions leading to their children being placed in special schools, believing that their intellect had not been reliably tested and that they themselves were not adequately informed of the consequences of consenting to their children being placed in special schools—they asked for the Education Authority to revoke the decisions made. The Education Authority declined, and applicants lodged a constitutional appeal on the basis of the *de facto* discrimination experienced. The Constitutional Court dismissed the relevant applicants' appeal.

Believing that they were treated less favourably than characteristically different persons in similar circumstances (without reasonable and objective justification), on the basis of this the applicants complained of an Article 14 violation in conjunction with Article 2 of *Protocol No. 1* of the *ECHR* (1952).

Firstly the Court determined whether there a case of indirect discrimination arose from the facts. As proving indirect discrimination can be particularly difficult, the Court reaffirmed that the applicants benefited from less strict evidential rules in the process. To this end, the Court allowed the use of statistical evidence in support of the applicant's case—the Court found using reliable and significant statistics in assessing the impact on a measure or practice on a group was sufficient for *prima facie* evidence. The previously mentioned statistical evidence was submitted to the Court, and consequently the burden of proof was shifted to the respondent State (which was unable to furnish contradictory statistics). The Court accepted that a disproportionately high number of Roma children were assigned to special schools, stating that (*D.H. and Others v. The Czech Republic,* [2007]):

> Despite being couched in neutral terms, the relevant statutory provisions therefore had considerably more impact in practice on Roma children than on non-Roma children and resulted in statistically disproportionate numbers of placements of the former in special schools.

Secondly, the Court needed to determine whether there was an objective or reasonable justification for the indirect discrimination at hand. The Court did not find objective or reasonable justification based on the flawed assessment procedure for Roma children, stating that (*D.H. and Others v. The Czech Republic,* [2007]):

> ...at the very least, there is a danger that the tests were biased and that the results were not analysed in the light of the particularities and special characteristics of the Roma children who sat them. In these circumstances, the tests in question cannot serve as justification for the impugned difference in treatment.

Additionally, the Court found that the parents of the children were not sufficiently informed of the consequences of consenting to the assignment of their children to special schools in order for them to be in a position to have accepted the discriminatory measures. Consequently, the Court

6.4.2.3 Affirmative Action

As indirect discrimination is prohibited by the *ECHR*, states are consequently required to consider the impact of measures and policies on protected and vulnerable groups—positive actions are again required including the adoption of special measures that facilitate the equal opportunity for enjoyment of rights; such measures may even result in discrimination that favours vulnerable groups or persons above other groups or persons in similar situations (Harris *et al.*, 2009, p. 611; European Union Agency for Fundamental Rights and Council of Europe, 2010, pp. 35–39). Such measures might be considered reverse discrimination or, preferably due to being less ambiguous or open to misinterpretation, affirmative action. Where such measures are taken, they are not considered a breach of Article 14 (Harris *et al.*, 2009, p. 611).[135]

---

upheld that there had been a violation of Article 14 in conjunction with Article 2 *of Protocol No. 1*.

[135] A useful example of such measures which would tend to treat persons in similar situations differently but is nonetheless acceptable is the case of *Stec and Others v. the United Kingdom* [2006]. This case dealt with three applicants who contested sex discrimination in the application of social benefit allowances (injury benefits and retirement allowances). The prime point of contention was different age limits applied to men and women (60 for women and 65 for men). The injury allowance was converted to a State pension for women and men at these respective ages.

In the Court's assessment, it noted that (*Stec and Others v. the United Kingdom,* [2006]):

> Article 14 does not prohibit a member State from treating groups differently in order to correct "factual inequalities" between them; indeed in certain circumstances a failure to attempt to correct inequality through different treatment may in itself give rise to a breach of the Article...

The Court accepted the clear differential treatment, however opined on the basis of justification that (*Stec and Others v. the United Kingdom,* [2006]):

> It would appear that the difference in treatment was adopted in order to mitigate financial inequality and hardship arising out of women's traditional unpaid role of caring for the family in the home rather than earning money in the workplace. At their origin, therefore, the differential pensionable ages were intended to correct "factual inequalities" between men and women and appear therefore to have been objectively justified under Article 14 of the Convention...

The Court noted that increased parity between men and women in the work force could justify an equalization of the pension age requirement, however it essentially granted the State a margin of appreciation in determining when it would be appropriate to implement such a change. It stated (*Stec and Others v. the United Kingdom,* [2006]):

> In the light of the original justification for the measure as correcting financial inequality between the sexes, the slowly evolving nature of the change in women's working lives, and in the absence of a common standard amongst the Contracting States..., the Court

6.4.3 Fiduciary Theory, Equality and Discrimination Revisited

Human rights are ultimately quite minimal claims towards the fulfilment of basic needs that facilitate a life of secure freedom (free of domination and instrumentalisation at the hands of state and non-state actors). They are not constructs of telic egalitarian thought in that they do permit some inequality, and perfect parity between all people is not exactly their goal—merely equal opportunity for the realisation of rights. The human rights of the vulnerable however do command more urgency and care in their protection, as the vulnerable can be disadvantaged in ways that render them more exposed to instrumentalisation and domination, they must surely be prioritised and adequately accounted for when the state takes action. The plight of particularly vulnerable groups has been well recognised by the international community, and the imperative to take special measures to protect their rights has been well enshrined in IHRL.

The practice of the ECtHR broadly complies with Fiduciary Theory and provides some instruction useful for the analysis that follows. In-keeping with the principle of formal moral equality, the ECtHR has prohibited different treatment of persons in similar situations without an objectively justifiable reason or proportionality in the means and goals being pursued. Discrimination however is not absolutely prohibited, it must merely not be arbitrary; that is, it must be reasonable and proportionate, much like standard human rights limitations.

Importantly, the practice of the ECtHR also exposes the problem of indirect discrimination, which highlights that while measures implemented by the state may be done without malevolent intent, the consequences can still disproportionately adversely affect vulnerable members of society. Any state that engages in measures which would tend towards violating the rights of its subjects, regardless of whether or not there was ill intent, strains the fiduciary relationship and the state's legitimacy as sovereign.

> finds that the United Kingdom cannot be criticised for not having started earlier on the road towards a single pensionable age.

In conclusion, the Court found no violation of Article 14, and found that the different treatment was justifiable given its historical purpose of ameliorating the situation of economic disadvantage faced by women (*Stec and Others v. the United Kingdom* [2006]):

> ...the Court finds that the difference in State pensionable age between men and women in the United Kingdom was originally intended to correct the disadvantaged economic position of women.

Situations of *de facto* discrimination can still place persons in positions where they cannot be said to be in secure and equal freedom under the rule of law. And the fiduciary cannot be said to be ruling effectively with solicitude for the rights and interests of its subjects when its action have drastically unequal outcomes that further marginalise the marginalised.

Importantly, ECtHR practice also underlines the importance of positive state action in ensuring that all have an equal opportunity to rights, and that discrimination can be justified on the basis that it helps persons who otherwise would not easily be able to realise their rights under ordinary circumstances. That is to say that, if persons are treated the same, despite having different circumstances, the fiduciary can be at fault, and should organise its measures in a manner that, if there is discrimination, it benefits the most marginalised.

As well as broadly complying with the principles of Fiduciary Theory and supplementing it with some form and content, the practice of the ECtHR provides important analytical principles for what follows. Practice of the ECtHR shows how discrimination can be assessed, by firstly identifying different treatment (admittedly, whether or not a pre-existing right has been violated is perhaps immaterial though it is a good starting point by offering context of discrimination), whether the situation of persons experiencing differing treatment is analogous, whether their personal circumstances are analogous, and whether the discrimination can be said to have had a legitimate aim and been proportional. It also indicates that statistical evidence can justifiably be used to support or assess charges of indirect discrimination (*D.H. and Others v. The Czech Republic,* [2007]).

6.4.4 Slándáil-type Systems and Discrimination

The final task here is to establish what the implications of Slándáil and similar systems might be for the broad principle of non-discrimination.

In terms of direct discrimination, such systems might be of limited but not negligible impact. Distinctions between persons could potentially be made as regards the right to privacy. The Slándáil system itself is designed in an inherently neutral way with regards to its impact on privacy vis-à-vis direct discrimination to the extent that it does not target persons based on particular characteristics, merely the content of their messages. This was explored extensively in the preceding chapter. This is not to say however that

the system, like any tool, cannot be misused by its end-users or service providers to some degree. In a worst case scenario, it would not be impossible for end-users to record, store, and disseminate information pertaining to individuals in a discriminatory manner. Consider for instance if such systems were deployed in states with deep racial divisions, characterised by mistrust between minorities and statutory agencies. In such a situation, emergency managers (possibly representing law enforcement) may treat information relating to Black and White individuals differently. Images retrieved by the system containing Black individuals, for instance, could be stored indefinitely without justifiable reason. Such difference of treatment between different persons in similar situations would be arbitrary, would not follow a legitimate aim, and could not be justified.

Similarly, where racial discrimination is prevalent in a society and institutionalised, emergency managers may choose to disregard filtered images that show ethnic minorities in distress, or ignore messages emerging from geographical areas known to contain a high population of ethnic minorities. In such cases, discrimination cutting across the rights of housing, property, and life would be possible results.

In both cases, such human rights violations would be facilitated in environments with rather entrenched institutional racism or discrimination, and the system itself would not be directly accountable for the rights violations[136]—nonetheless, whilst the system is ostensibly neutral, it would be a tool complicit in such human rights violations, and would enable them, regardless of whether or not it was the only tool available on which to base discriminatory actions or measures.

In both highlighted cases, no derogation or rights limitation would be applicable to the discriminatory practice.

The more likely scenario to occur is use of the system contributing to indirect discrimination (at least where emergency managers are excessively dependent on Slándáil or similar systems for situational awareness). Social media posts may emerge asymmetrically from different geographical areas dependent on their socio-economic demographics and therefore bias the pool of information available to emergency managers in favour of particular groups, typically the youngest and most educated. In

[136] Which would theoretically be possible with or without such a system, given that emergency managers could base discriminatory decisions on a plurality of pre-existing or incoming information.

states where there are significant gaps between men and women either by internet access, or literacy, and where consequently men may be disproportionately represented on social media, men's needs and interests may too be disproportionately represented and result in disaster response that favours the fulfilment of those needs. In states with particularly poor rates of internet users (for instance, Sub-Saharan African countries) and internet access, the most privileged in society (likely young, educated, and wealthy) will be dominantly represented on social media, and therefore disaster response would be vulnerable to being biased in their favour. Where states have large populations that do not speak the dominant language, and where such languages are not accommodated by systems that harvest data from social media (recall that at time of writing, Slándáil is only operational in English, German, and Italian), the needs of such populations may be well represented but nonetheless filtered out with the "noise" by the system. Categories of disadvantage vis-à-vis social media use and natural disaster vulnerability may also overlap significantly, and whereby minorities face multiple overlapping disadvantages they may be particularly vulnerable to the impacts of indirect discrimination. This is not an exhaustive list of possible examples.[137]

Persons with various personal characteristics, as explored in section 6.2, are acutely vulnerable to natural disaster, and evidence has borne that in the existing situation they are victims to indirect discrimination in response and recovery (the evidence being their worse natural disaster outcomes). To view discrimination as the ECtHR does, in conjunction with other rights, the fulfilment of their rights to life, shelter, and housing, may not be on equal terms with more privileged members of society. The introduction of technologies that harvest data from social media during natural disasters, with regards to the potentially uneven representation of vulnerable groups on social media, threatens to compound this indirect discrimination by biasing the pool of available actionable information in favour of privileged groups (where emergency managers depend excessively on social media as a source of information).

Indirect discrimination in emergency response (and recovery) to natural disaster is emblematic of general state failure to adequately protect the rights of society's most vulnerable, and of a failure in planning and preparation to adequately mitigate the

[137] Consider Irish Travellers, who have lower literacy rates, who live with low employment, education, and in poor health in disproportionately large numbers, as well as frequently living in accommodations—mobile homes—more exposed to the elements[137] (Kelly *et al.*, 2014; Sheeran, 2015).

impacts of natural disasters on the vulnerable. This is evidently one of the challenges faced by disaster management at the moment, and is not likely a product of intent so much as a symptom of ignorance. Systems such as Slándáil have the potential to shift attention further away from the vulnerable, however in such cases the issue emerges more from a failure of planning for the needs of the vulnerable than the use of Slándail-type systems—it produces information that needs to be weighed in context with other information. Slándáil itself may enable indirect discrimination, however only if it is misused or used poorly, or in a vacuum of complementary information. It does not decide how response should unfold, merely provides information to assist this response.

The risk of social media data dominating decisions in disaster response is currently still low. As stated earlier, emergency managers rely on a plurality of information sources, and therefore should not be basing decisions exclusively on information brought to their attention by systems such as Slándáil. To that end, Slándáil may not necessarily make an evidently poor situation of indirect discrimination (globally) worse.

Statistical evidence going forward may help determine the true impact of such systems on emergency response. It may well be necessary to conduct case studies in the future that compare the efficacy of emergency response on similar pre and post Slándáil-type system managed disasters, in order to determine if such systems do actually contribute to emergency response that favours the privileged (even more). At that point, it will be easier to identify problematic aspects of emergency management, and the use of such systems, that systematically disadvantage the vulnerable.

It remains important that indirect discrimination be eliminated, with or without the addition of Slándáil-type systems to the overall situation, particularly with regard to the fiduciary duty to provide a regime of secure and equal freedom, even if that means implementing measures that favour the vulnerable more than those already particularly resilient to instrumentalisation and domination. As stated earlier, whilst having the theoretical capacity to contribute to circumstances that disfavour the vulnerable, systems such as Slándáil can be built, or adjusted, to either mitigate this threat or even offer information that mitigates indirect discrimination in disaster response on the whole. Recall that a variety of geo-spatial statistics can be loaded onto interactive maps (through SIGE or similar GIS systems), and that computational solutions can be applied to model combinations of datasets to indicate vulnerability.

It is important, with regard to the reality of indirect discrimination, and the capacity of Slándáil-type systems to compound the problem, that they are designed to present emergency managers with as full a set of information as possible, not just social media activity, but the locations and risks to the most vulnerable. It is also important that such systems are designed to recognise as many languages as possible that are used in their states of operation. Software designers then have responsibility, by virtue of the authority delegated to them in providing these services that means that they partake in the fiduciary relationship between statutory authorities and subjects, to ensure that their systems function in a manner that can aid the best human rights outcomes, with minimal adverse human rights impacts.

The use of the Slándáil system, and similar systems, certainly follows a legitimate aim (protection of life and property), though there is a question mark over the proportionality of its implications for human rights. At worst, where institutional practice is so weak that reliance on social media data trumps all other information sources, it may disproportionally benefit those already privileged in society, it may treat all alike *despite their differing circumstances* and would contribute to arbitrary human rights interferences. At best, where the system is adequately designed to convey a plurality of information itself, and where a plurality of information is accounted for, and where needs based assessment of populations in disaster is thoroughly accounted for, it can deliver information that informs and supports a fairhanded, proportional response—if not response that prioritises the needs of the most vulnerable.

6.5 Conclusion

This chapter contributed to the disclosive analysis by interrogating the ethical and human rights implications of design choices and potential uses of Slándáil-type systems.

From an ethical perspective, with regards to global inequality—which may vary in severity by region but is always nonetheless inescapable—compounded by the digital divide, it is evident that the introduction of Slándáil-type systems in natural disaster management poses some significant threats the equitable emergency response. Those in vulnerable positions, with few functionings and challenged capabilities, already stand to be the worst affected by natural disasters. In a worst case deployment scenario, systems such as Slándáil could stand to reinforce the negative experiences of vulnerable populations in disaster, where they may essentially be shut out of the empowered collective action described in Chapter 4 where they are unable to engage meaningfully

with social media. Slándáil-type systems can only cast a reflection of those who hold the mirror; the emergency manager cannot see—using a Slándáil-type platform—these reflections where there is no mirror to hold. Vulnerable voices may be lost in the din. Worse yet, malicious emergency managers may even intentionally ignore the plight of vulnerable populations. From the preceding analysis, it is clear that such a cases are an injustice, society as a whole has a responsibility towards its most vulnerable constituents, and where they do not have the agency to help themselves the absent functionings that make this so ought to be provided, or compensated for. To that end, Slándáil-type systems should not be depended on too much for their social media harvesting capabilities. At the same time, whilst they can contribute to injustice, such systems do have open-ended functionality and the ability to display a variety of datasets to emergency managers regarding the locations of vulnerable populations,[138] or computational artefacts can be designed to actively calculate and weigh combined datasets to establish risk indices.

Ultimately, the real challenge extends beyond this particular context. Responsible MASes (particularly governments), must ensure that emergency management is inclusive and sensitive to the particular needs of vulnerable populations at all phases of emergency management—and these needs should be prioritised.

A human rights analysis of the situation does not lead to radically different results. Such systems can be used to discriminate in a direct manner (a point which will briefly be returned to in the following chapter) and intuitively can lead to a high risk of indirect discrimination where emergency managers rely too heavily on information gleaned from social media, potentially resulting in emergency response stilted in favour of social media users. The state, duty-bound to provide a regime of secure and equal freedom, must strive to avoid such situations where some of its subjects are so vulnerable to the impacts of disaster, and must mitigate any factors that compound this insecurity, even if that means that it must enact policies that disproportionately benefit its more vulnerable subjects—that is, again, the vulnerable should be prioritised in emergency management.

When designing Slándáil-type systems, the ethical developer ought to inscribe these systems with functionality that mitigate the possibility of emergency response being

[138] Both socio-economically and physically with regards to natural hazard proximity as the case may be.

biased in favour of the privileged. Emergency managers ought to ensure that they are using a plurality of information sources, and are paying particular attention to the needs of the vulnerable.

The discussion here allowed for an understanding of the potential issues concerning (in)justice in emergency response to natural disaster vis-à-vis Slándáil-type systems, and in uncovering the risks some solutions became apparent. The task of proposing mitigating measures is not yet complete, but will be returned to in Chapter 9.

On a final note is that one lesson in particular became very apparent in this Chapter, which is, emergency managers cannot allow information derived from social media to exclusively lead emergency response.

7 TRUST

7.1 Introduction

In this chapter, the implications of Slándáil-type systems for the value of trust will be examined. Trust is an important value in society, allowing people to overcome uncertainty and place some belief in others to act towards certain shared goals (Nissenbaum, 2001; Taddeo, 2009; McLeod, 2015). Trust is something then that intuitively generates some efficiency and fosters co-operation in society, empowering it to operate more effectively on the whole. In natural disaster situations in particular one can expect to place trust in a variety of actors; the media for accurate information, members of the community for help and comfort, and of course emergency management and response agencies for relief and assistance. The introduction of social media and autonomous artefacts that scan and process information therein adds a new dimension to trust-qualified relations in natural disasters, with numerous implications that will be analysed in what follows.

The following will introduce the reader to standard challenges to trust emerging from social media and technology generally, with particular reference to misinformation/disinformation and the possibility of function creep.

It will outline traditional theories of trust and justify a more contemporary interpretation that can be extended into the digital space and include artificial agents. Following this, this theory of trust will be applied to examine the interactions of human and artificial agents in natural disaster response, and the factors that undermine trust between agents. The analysis here will move tentatively beyond the theme of natural disasters, accepting that Slándáil-type systems can evolve to act beyond such scenarios, and such an evolution of functionality has implications for trust-qualified relations.

This chapter will advance analysis to the domain of human rights under Fiduciary Theory, arguing that the public trust is undermined by violations of rights, thereby opening up the discussion to a number of rights which, following ethical analysis, were selected based on their relation to the potential chilling effects emerging from the utilisation of Slándáil-type systems. The rights selected are the freedom of expression, association, and assembly.

7.2 Social Media, Digital Technology, and Challenges to Trust

7.2.1 Misinformation and Disinformation

The credibility of information encountered on social media during crises will naturally have implications for trust relationships, with implications for the trustworthiness of information sources, those acting on social media derived information (emergency managers and responders), and in this particular case the mediator between the two that is delegated the task of filtering potentially actionable information (EMIS such as Slándáil). False information[139] intuitively emerging from un(der)informed or malicious agents has the potential to lead to inappropriate action that can damage trust relations within multi-agent systems, potentially causing mutual distrust—consider emergency managers who deploy assets based on erroneous or misleading reports, and then the general public which may lose trust in emergency management agencies to use their powers responsibly. These trust qualified relationships and potential impacts to them will be analysed in greater detail presently. This subsection will briefly explain the dynamics of rumour spread on social media during emergencies, where rumours broadly represent misinformation[140] and disinformation.[141]

According to Ozturk, Li, and Sakamoto (2015, p. 2406), social media provide a "...rich substrate for rumour propagation," and when such rumours gain traction, propagating from person to person through retweets or shares, they take on a viral pattern, and spread like an epidemic—sometimes even passing the threshold from social to legacy media (Farhi, 2012; Takayasu *et al.*, 2015, p. 34). Apropos to this research in particular, some describe the phenomenon of the online propagation of rumours as *digital wildfires* (World Economic Forum, 2013).

A wealth of research has been conducted examining this phenomenon, and whilst not new, it has been greatly enhanced by the capabilities of digital technologies that enable rapid dissemination to wide audiences (Mendoza, Poblete and Castillo, 2010; Castillo, Mendoza and Poblete, 2011; Seo, Mohapatra and Abdelzaher, 2012; Yang *et al.*, 2012; Gupta *et al.*, 2013; Oh, Agrawal and Rao, 2013; Starbird *et al.*, 2014).

The potential risk of false information contaminating the pool of information that social media provides should not be underestimated, and without due diligence being taken in

[139] Which technically does not truly qualify as semantic information (Floridi, 2011b).
[140] False information spread irrespective of intent.
[141] False information spread with deliberate intent to deceive.

confirming information there is likewise potential for great harm; rumours can "...increase the sense of chaos and insecurity in the local population" (Castillo, Mendoza and Poblete, 2011, p. 5).[142]

Rumour spreading can occur not simply as textual information, but can also utilise doctored or re-contextualised photographic media (Gupta *et al.*, 2013). Figure 20, for instance, shows an image that was shared as legitimate during Hurricane Sandy, but is in fact production art from the movie "The Day After Tomorrow" (Farhi, 2012).

Figure 20: Fake image circulated on social media during Hurricane Sandy (Source: Farhi, 2012)

Whilst no consulted research indicates that rumours led to misallocation of resources by emergency management professionals, the online propagation of rumour has been demonstrated to cause real life harm. One example is the case of rumours arising from digital vigilantism in the aftermath of the Boston Bombing, where digital vigilantes on reddit.com falsely identified a missing *(innocent)* student, Sunil Tripathi (who was later

[142] The volume of persons involved in rumour diffusion can be quite high in absolute terms, research by Takayasu *et al.* (2015, p. 3) focused on rumour diffusion and convergence during the Japanese 3.11 earthquake showed that in one instance 38,226 twitter users were involved in rumour diffusion (originating from a twitter user with 360 followers) during that event. The rumour relating to a gas leak that occurred in the aftermath of the earthquake, which was "Please spread: To those who live close to the east shore of Tokyo Bay! Due to the explosion of oil tanks, harmful chemical materials may fall with rain soon. Bring your umbrella and rain coat with you to protect your skin from dangerous rain!!"

A simulation by Ozturk, Li, and Sakamoto (2015, p. 2410) demonstrated that a Twitter rumour could reach an audience of 60,000 persons through 132 initial posters.

found dead) as an assailant, thereby causing distress to his surviving family (Suebsaeng, 2013; Starbird *et al.*, 2014; Piven, 2016).[143]

In attempting to explain why rumours may gain such traction over social media during crises, Ozturk, Li and Sakamoto (2015, p. 2406) offer the following explanation, "...online, the ability to participate pseudonomynously, low levels of entry barrier, social presence, and lack of gatekeeping mechanisms existing in traditional mainstream media create a setting of low accountability and uncertainty." In research that analyses the reasons for rumour generation in social crises in more depth, Oh, Agrawal, and Rao (2013) found that factors such as source ambiguity,[144] and (to a more marginal extent), personal involvement were relevant in rumour generation.

Much of the research consulted indicates that rumours are countered, indeed, often counter-rumour or corrective posts will outweigh rumour exclusive messages, especially where official statements or sources exist to dispel those rumours (therefore source ambiguity is mitigated)(Mendoza, Poblete and Castillo, 2010; Ozturk, Li and Sakamoto, 2015; Takayasu *et al.*, 2015)—though the "correction signal" can lag behind rumour spread (Starbird *et al.*, 2014, p. 661; Takayasu *et al.*, 2015, p. 3).

In sum, with regard to the open, largely unregulated nature of social media, in combination with situations of anxiety and little official information (exacerbated by the potential for malicious agents that can act anonymously with little accountability), untrustworthy information can pollute the pool of potentially actionable information, which is spread by misinformed or simply untrustworthy individuals. The presence of untrustworthy information on social media may strain trust between emergency manager and citizen, and intuitively inappropriate actions made based on untrustworthy information may undermine the citizen's trust in the emergency manager. The failure of the mediating tool (Slándáil-type systems) to produce trustworthy information could result in a loss of trust of both agents in such systems.

[143] In this era where "fake news" is becoming a hot topic, the true capacity of digital wildfires to cause harm outside of cyberspace is beginning to make itself apparent, and warrants further research, however it is beyond the scope of this research to investigate it further (World Economic Forum, 2013; Hunt, 2016).

[144] Where the source of information being reported is ambiguous.

7.2.2 Function Creep

Another issue with implications for trust relationships in a MAS consisting of emergency managers, citizens, Sládáil-type systems and social media is that of function creep. This particular challenge is a complex one and its potential for adverse impacts is contingent on a number of variables. Here, the potential for function creep to erode trust qualified relationships will briefly be described before again returning to a deeper analysis later on.

Function creep broadly refers to a situation where a procedure (or technology) implemented towards one specified goal or purpose is gradually (or perhaps even quite rapidly, as the case may be) implemented towards a wider set of goals or purposes (O'Brien, 2008, p. 31; Dahl and Sætnan, 2009; Backman, 2012, p. 277).[145]

Such a process of widening scope of use may not represent an inherently unethical problem nor necessarily compromise trust, it may simply represent natural evolution, particularly in shifting moral landscapes. A good case in point is the example of criminal record checks in Sweden, which implicates both procedures (mandatory disclosures of criminal convictions) and technology (centralised criminal databases). In Sweden, following multiple sex crime scandals involving minors, the Government moved towards mandatory criminal background checks for applicants of jobs in schools and preschools (initially), whereby applicants were required by law to disclose their criminal records (Backman, 2012). This was in the initial case an example of function creep as such procedures evolved from the protection of national security to the protection of children (Backman, 2012, p. 277). This evolved still to include the licensing of medical practitioners by the National Board of Health and Welfare in the wake of a scandal involving a murder convict being accepted into medical school (Backman, 2012, p. 284). Such examples of function creep are debatably acceptable, existing data is utilised towards the protection of vulnerable persons in a relatively discriminating and controlled manner, and the society itself, in the wake of specific moral outrages,

[145] Explaining why function creep occurs, Dahl and Sætnan (2009,p. 88-89) argue that:

> ...we often see technologies introduced when conditions are taken to indicate a dire need, then gradually expanded into less urgent uses... It may also come about in spite of everyone's best intentions, for instance through uncritical optimism and because the moral terrain shifts as soon as the initial investment is made. Once a technology is in place, it becomes wasteful not to use it to the fullest acceptable limit. Usages that might not have been sufficiently legitimate for initial implementation do have sufficient legitimacy to be tacked on later.

reached some agreement (these new procedures were effectively normalised) that the value of privacy could acceptably be re-weighted in the pursuit of other values (the personal security of the vulnerable)(Backman, 2012).

In other cases the function creep may be more insidious in consequence even if not by intent, and not so much an evolution of systems and procedures but a problematic mutation that may both expose or exacerbate existing distrust and undermine trust in multi-agent systems. As has been illustrated in Chapter 4, social media provides a potential treasure trove of information for emergency managers, but such a trove is rich in personal data, and additionally this trove of data can be processed in any number of ways that go above and beyond the particular interest of this research, emergency management, with other potential uses being crime prevention and national security. Social media provide a fertile ground for surveillance, with social media users acting as effective sensors of their environment and events occurring within it. Indeed, the applicability for analysis and inspection of social media data is wide, posing uses such as community tension monitoring on something of a more macro-scale, to providing access to private communications for the purposes of the protection of national security (Ball, 2013; Williams *et al.*, 2013).[146]

Though not an example of function creep pertaining to social media data, the example of forensic databases provides a useful illustrative example of function creep which may potentially exceed ethical boundaries,[147] or at least cause grounds for discomfort and distrust. In some cases, familial DNA searching has been used on DNA databases during investigation activities in order to identify criminal suspects where DNA samples have been found at a crime scene but no match has been found in the DNA database (Bhattacharya, 2004; Dahl and Sætnan, 2009). Dahl and Sætnan (2009, p. 92) warn in this case that "[t]his opens not only for surveillance of convicts (or even suspects) but also their families."

Surveillance of wide and/or increasing scope is a threat to privacy, under particular circumstances, and can be in itself unethical if it unjustifiably interferes with privacy,[148]

[146] The purpose of this research is not to assess the legality or ethics of these two examples, they simply serve to highlight the diverse possibilities for social media data analysis

[147] An ethical analysis of this practice will be set aside here, however the practice, involving privacy intrusion of innocent individuals, is sufficiently dubious to serve as an adequate example of practices that may shake trust.

[148] As has been argued, where the flow of personal information is inappropriate and no sufficiently compelling grounds can be offered for norm deviation.

and in such cases can intuitively undermine the public trust in statutory agencies and the technologies they have at their disposal. Even where such violations can be argued to be justifiable, they may plausibly cause a discomfort that nonetheless undermines public trust. A consequence of undermined trust can be uncertainty regarding the force of norms, an uncertainty that can have a chilling effect on persons who are arguably unsure, essentially, of whether there will be repercussions for their innocuous or otherwise legitimate actions (Fox, 2001; Penney, 2016; Warner and Sloan, 2016).[149] This chilling effect can be argued to be either an example of trust eroding with the introduction of surveillance methods causing uncertainty, or be an expression of pre-existing distrust in governments—for the time-being it is unnecessary to arrive at a conclusion, as in either case it represents a manifestation of distrust triggered by state action.

A prime example of this chilling effect of surveillance, particular surveillance with arguably ambiguous parameters, is rendered in the research of Jonathon Penney (2016). Penney's (2016) research follows the volume of Wikipedia searches of terrorism related articles after the Edward Snowden disclosures of the NSA's advanced online surveillance capabilities. Penney (2016, pp. 147–161) found a significant and ongoing reduction of views of the selected articles occurring immediately after the Snowden disclosures in his dataset (though November 2014 was an outlier, coinciding with an Israeli Military offensive on the Gaza Strip—this buck in trend was considered an "exogenous shock").

While widening or even illicit surveillance by the state targeting the public has the potential to undermine trust of the public in the state, the nature of social media, particularly combined with the gaze of the state, also has the potential to undermine trust between members of the public.

Social media enables persons to, in essence, surveil each-other,[150] in what can be called participatory, lateral, or inter surveillance—citizens may cast the investigative gaze on each-other and can disseminate their captured information on social media, and, where there is a statutory body watching over the sphere of social media, an image of an Orwellian system of citizen informers is cast (Jansson, 2012; Purenne and Palierse,

[149] That is, whether the state will turn its surveillance on them, or indeed whether consequences of its surveillance may result in some censure for them (Fox, 2001; Penney, 2016; Warner and Sloan, 2016).

[150] To elaborate on this, to watch or follow the social media accounts of their peers, or, outside of the digital space, to record and watch others and share this information on social media which was generated in the 'real world.'

2017). Lateral surveillance, in its traditional, analogue sense in the community-watch form has had mixed impacts, which have not altogether been negative nor do they categorically cause distrust, however some evidence has shown it to, at worst, reinforce prejudices and suspicions (and ultimately distrust) concerning the "other", or strangers and the marginalised from outside of the community (Purenne and Palierse, 2017, pp. 90–91).

Everything posted online can be scrutinised and discussed in public and matters of law enforcement can be devolved to the private citizen, potentially with negative consequences. When injecting legitimate law enforcement into this sphere, who could potentially be fed inaccurate information, worst case scenarios could become even more severe (perhaps arrests based on poor community intelligence). The potential negative consequences of interveillance give rise to reason for distrust between citizens, distrust of citizens by the state agents, and even distrust in mediating systems such as Slándáil.

Research has shown that the advent of social media has had chilling effects that influence behaviour outside of the realm of cyberspace, that individuals moderate their behaviour in the real world, aware of the presence of recording devices and the possibility that their business will be shared in the online sphere of social media (Marder *et al.*, 2016).[151] Such research suggests that individuals moderate their behaviour for fear of censure or judgement, or the unpredictability of consequences of sharing information relating to their lives that they do not want widely shared, and thus hinting at the possibility of a lack of trust in their peers, that is, a lack of trust not to be judged or censured by them.

In sum, function creep in the context of systems for monitoring social media can support widening surveillance, and such surveillance, at worst, might lead to a chilling effect whereby persons moderate their personal behaviour for fear of censure—this may either expose existing distrust, create distrust, or even give reason to cause distrust in multi-agent systems, mutually between all comprising (human) agents.

[151] Recall the girl with pink hair, who certainly would have been given cause to question her every move in public life upon discovery of the "Madrid" Flickr group.

7.3 Information Ethics and Trust

7.3.1 Trust and e-Trust

Traditionally trust, in its most elemental form, was most succinctly and effectively defined by Annette Baier (1986, p. 236) as a three-place predicate whereby A (the truster) trusts B (the trusted) with some valued thing C. It must however be distinguished from mere reliance; trust is not something that is simply disappointed when C is not executed properly, it is betrayed, it entails vulnerability and exposure—it may entail risk to some harm (Baier, 1986; McLeod, 2015). In trust relationships, as typically formulated, the trusted must have sufficient autonomy to not execute C; if constraints are attached to the trusted that limit this autonomy and therefore their capacity to betray, where the trusted does not have some form of discretionary vulnerability and the truster limited vulnerability, it might be that the truster does not indeed trust the potentially trusted (McLeod, 2015).

Baier (1986, p. 234) adds that trust renders us vulnerable to the *good* or *ill will* of others. Indeed, mental states are implicated in trust relations; care for a truster, or optimism or pessimism regarding a trusted's trustworthiness or ability to execute C are factors in trusting—trust can be a cognitive or non-cognitive (or affective, such as with the infant child and their guardians) exercise (Baier, 1986; Taddeo, 2009; Ess, 2010; McLeod, 2015).

Discrimination should be offered in what or whom one trusts, and it should not be placed arbitrarily, because trust implies some risk of betrayal—trust should therefore only be given where it is "warranted", or "justified or well grounded" and "plausible" (McLeod, 2015). Too much (or blind) trust increases vulnerability, whereas when trust is justified or well grounded, and plausible, the risks of trust are reduced (McLeod, 2015). Nissenbaum (2001, pp. 110–112) expands on this point, offering several criteria by which trust is deemed appropriate, or merited, including; History and Reputation; Inferences Based on Personal Characteristics; Relationships: Mutuality and Reciprocity; Role Fulfilment; and Contextual Factors.[152]

[152] To elaborate on Nissenbaum's (2001, pp. 110-112) criteria:

- History and reputation — the history of interactions between one person and another enables an assessment of the trustworthiness of the potentially trusted. Where no personal history exists, the truster may refer to the reputation of the potential trusted.

The contextual factors element of trust grounding is similar also to theories of social constraints providing the basis for trust, that is, trust can be justified on the basis that social (or indeed legal and political) forces require an actor to act in good faith as the threat of sanction hangs over them should they deviate from whatever social norms are implicated (McLeod, 2015). Whether someone can be trustworthy on the basis that they can only be trusted on the fear of sanction however is debatable, as someone is arguably not truly trustworthy, merely predictable and reliable and capable only of disappointing and not betrayal in this scenario (McLeod, 2015).

Trust is, by most traditional accounts, a very human condition that can only be experienced by human agents. Trust requires rational assessments and risk-taking, vulnerability and exposure to the discretion of others and their intentions, it may entail attitudes, care, and good will. The introduction of the digital space, and artificial agents, complicates trust and trust giving.

- Inferences based on personal characteristics — trust is granted where shared values or experiences are perceived, and/or where the potentially trusted presents characteristics of "...virtue, loyalty, prudence and a desire for the good opinion of others...".
- Relationships: mutuality and reciprocity —the nature or structure of a relationship may inform whether there is trust. Nissenbaum presents the example of "common ends" relationships, where both parties are in similar situations with the same ends, and one has responsibility over the other. Nissenbaum offers the example of the trusted pilot, arguing "...I place trust in the pilot partly because he is in the plane with me and I presume that we have common, or confluent, ends; our fates are entwined for the few hours during which we fly together." Reciprocity provides grounds for trust in what Nissebaum refers to as "tit-for-tat" exchanges where there are no common ends but the truster may, for example, commit some action in the expectation that in reversed positions the trusted will do the same, specifically or more generally.
- Role fulfilment — role fulfilment provides grounds for trust where the trusted warrant trust because of the nature of their particular role, and their assumed commitment and responsibility towards executing their tasks. The truster may know the background and parameters of the role and can extend trust based on this knowledge, as Nissenbaum explains, referring again to the pilot:

  > Crucial to my trusting the pilot is that he is a pilot, and being a pilot within the framework of a familiar system has well articulated meaning. I know what pilots are supposed to do, I am aware of the rigorous training they undergo, the stringent requirements for accreditation, and the status of airlines within a larger social, political and legal system.

- Contextual factors — contexts, or settings ranging from families to towns, affect readiness to trust. This criterion is more complex than the others, however suffice it to say settings of transparency, where betrayals are punished, trust promoting norms are promulgated (through culture, for instance), and where trust can essentially be insured, for example, "such a policy is the current arrangement of liability for credit card fraud, which must surely increase people's willingness to engage in credit transactions."

The introduction of the digital environment, or the wider infosphere as it exists as the sum of Being, and the constituting interactions between entities, both human and artificial, poses challenges for the older conceptualisations of trust, and its implementation in more ambiguous, if not obscure, contexts. The online space, inhabited by human and artificial agents, poses challenges or obstacles to trust, and it can be difficult for agents (human or artificial) to establish the necessary grounding to place trust. The online space is missing the cues that give rise to trusting attitudes; anonymity can hinder trust as agents can be unknown and there may not be a past history of experience with others (reputation) to refer to, personal characteristics are obscured that can make the presence of shared values difficult to discern, and contexts can be inscrutable—constraints to betrayal provided by setting (such as norms) are unclear (Nissenbaum, 2001b, pp. 113–114). Social and environmental cues then that facilitate trust may not be, or be weakly, present in the online space, and conditions that facilitate distrust may even be stronger, for instance obscured identity and a lack of accountability. The online environment then can be one of great uncertainty where typical formulations of trust seem either inapplicable or weakly applicable.

Without modifying our understanding of trust under these conditions, it can be difficult to argue that trust can exist at all between two agents as mediated by electronic environments. And yet it must exist, as demonstrated by persons willing to make near anonymous sales transactions online, or interacting (perhaps even intimately) with others over forums and online chatting services (Turilli, Vaccaro and Taddeo, 2010). These people, digital natives and immigrants, are not simply gullible, they are basing their trust assessments on other factors perhaps not always present in standard trust justifications.

The traditional conception of trust, interpersonal or as referred to as face-to-face (or f2f) trust by Grodzinsky, Miller and Wolf (2011, p. 17), predominately describes trust between humans in physical environments. Trust now however must be able to describe apparently trusting relations between human and artificial entities both inside of and outside of cyberspace. Grodzinsky, Miller and Wolf (2011, p. 25) introduce several subclasses of trust relations that demonstrate just how complex trusting relations can be in the modern infosphere:

- HHP-trust: traditional notion of human, "face-to-face" trust
- HHE-trust: humans trust each other, but mediated by electronic means

- HAP-trust: humans trust physically present AA, for example, a robot (no electronic mediation)
- HAE-trust: human trusts an artificial entity (like a web bot) over the internet
- AHP-trust: an AA, perhaps a robot, trusts a physically present human
- AHE-trust: an AA, perhaps a web bot, trusts a human based on Internet interactions
- AAP-trust; an AA trusts another AA in a physical encounter; because this is P [physical] trust, the AAs might, for example, use sign language
- AAE-trust: an AA trusts another AA electronically, e.g., two web bots communicate via the Internet

What these sub-classes of trust show is that trust is not necessarily the exclusive domain of the human agent, that a satisfactory theory of trust must be able to account for trust between non-human agents or entities, and therefore that notions of trust and trustworthiness, or justifications and grounding of trust, must extend beyond notions of good will, caring, and mental states. Secondly, these sub-classes broadly fall under two parent classes of trust; f2f trust, and eTrust, where the relationship is mediated by electronic environments.

An exploration of eTrust, and the reflection it demands on trust generally, can shape a modified definition of trust that can be deployed for analysis presently. The work of Mariarosaria Taddeo, Matteo Turilli and Antonino Vaccaro provides excellent assistance is clarifying the role of trust as it operates across physical and digital space (Taddeo, 2009, 2010b; Turilli, Vaccaro and Taddeo, 2010). Taddeo's (2010b, p. 247) definition of eTrust builds on her definition of trustworthiness with regards to AAs in distributed systems, which is "...understood as a measure that indicates to the trustor the probability of her gaining by the trustee's performance and, conversely, the risk to her that the trustee will not act as she expects." At base then, an agent is trustworthy where the rationally assessed likelihood of the trusted in successfully executing its entrusted action is higher than the probability of it deviating from expected performance. Trustworthiness in this sense is not general but based on an AA's ability to work towards specified goals (so the AA may be trustworthy in executing one action, but not another) (Taddeo, 2010b, p. 248).

Following this, Taddeo's (2010b, p. 255) definition of eTrust (at a high level of abstraction) is:

> Assume a set of first order-relations functional to the achievement of a goal and that at least two agents (AAs or HAs) are involved in the relations, such that one of them (the trustor) has to achieve the given goal and the other (the trustee) is

> able to perform some actions in order to achieve that goal. If the trustor chooses to achieve its goal by the action performed by the trustee, and if the trustor considers the trustee a trustworthy agent, then the relation has the property of being advantageous for the trustor. Such a property is a second-order property that affects first-order relations taking place between AAs, and is called trust.

This definition leaves open the criteria for trustworthiness (Taddeo, 2010b, p. 255), which means that they may be selected as appropriate for whatever context to which one refers—therefore the previous criteria often associated with f2f trust need not be abandoned, but applied judiciously depending on the agents concerned and the context (Taddeo, 2010b).

In this definition of trust, trust is a second order property of first order relations, and it differs from other definitions of trust in this way (Taddeo, 2010b, p. 254). The first order relation might be some transaction whereby A delegates to B some valued C (with the goal being the completion or execution of C), and the second-order property is trust. A in this case needs to achieve C, whilst B has been delegated tasks to achieve this goal. This is a first-order relation qualified by trust, that is, "...if the trustor considers the trustee a trustworthy agent and hence does not supervise the trustee's performances, then the relation has the property of being advantageous to the trustor," and "[s]uch a property that affects the first order relations occurring between agents is called trust" (Turilli, Vaccaro and Taddeo, 2010, p. 340).

Whilst this approach to trust was formulated to explain occurrences of trust involving artificial agents, it is applicable to human agents too (Taddeo, 2010b, p. 254). This is achieved through abstraction and, again, no direct guidance on criteria for trust assessment is specified; it can be chosen on a case-by-case basis (Taddeo, 2010b, p. 254).[153]

In using this definition of trust, it can be argued and demonstrated that trust can be placed across online environments (and with justification). Formerly necessary criteria including shared values and certainty of identity are rejected as an absolute requirement (Turilli, Vaccaro and Taddeo, 2010). It is important that some relevant criteria are selected, though the presence or absence of identity and shared values are not

[153] The definition provided is of e-Trust at a high level of abstraction, which for the purposes of this research will be sufficient for analysis of trust qualified relations between human and artificial agents, or between human agents as mediated by digital spaces.

necessary. On the point of obscure identities, Turilli *et al.* (2010, p.337) reject that this is a major obstacle—they argue that "...it is true that identification, authorisation and accounting procedures are often in place so that, if needed, who interacts in a given online environment can be univocally identified," and that beyond this, referral systems (such as seller reviews on Amazon and Ebay) provides reputation transparency on which persons can factor their trust assessments.

The presence of care, or good will, in trust relations may also be widely unrealistic and unnecessary criteria for justification or defining trust. These would be unreasonable or unrealistic criteria as they would require an omnipotent ability from the trusting agent to assess the true intentionality of the trusted agent. These are subjective rather than objective criteria, and while one may be wise to place trust in those who overtly seem to care about them, or have good will generally, such appearances can of course be faked.[154]

In terms of shared values and norms, Turilli *et al.* (2010, p. 337) argue that as trust requires (by definition) risk, the assurance offered by values and norms replaces trust and that "... it is often assurance and not trust that fails to emerge in unstructured environments." Assurance however does make trust easier, even with constraining moral and social forces risk remains,[155] as while trust requires risk, factors that mitigate risk make it easier and sensible to extend trust. Therefore whilst the presence of shared values, norms and other such constraining forces should perhaps not be considered an absolute requirement for the grounding of trust or the emergence of trust, they are certainly rational criteria.

All this is to say that the criteria for justification for trust, and its plausibility in the physical and digital space, are dynamic and not absolute.

It remains to ask what distinguishes trust from reliance under this revision of trust? Disappointment and betrayal are affective terms with primary application to human agents. Both of these words imply different levels of severity, where being disappointed is simply being let down, and betrayed implies something more visceral, like breaking a promise or being exposed to danger from the treacherous revelation of information to

[154] Frankly, in many transactions one cannot truly expect that the potentially trusted cares at all or even bears good will, and even if they did not, it would not necessarily preclude them from being trustworthy unless they were actively malicious.

[155] One can be stabbed randomly on the street despite the assurance that the assailant will be prosecuted when caught—assurance does not absolutely eliminate deviant behaviour.

an enemy (Mirriam-Webster, no date). The difference is harm, or exposure thereof—betrayed expectations may be more harmful and dangerous than mere disappointed expectations. Whether one is disappointed or betrayed will vary by context. AAs cannot be disappointed, but perhaps can be betrayed as a function of harm; their own reputations are harmed whereby they trust other agents in the delegation of some action that the trusted agent fails to execute—the trusting AA may be decommissioned as a result. When an AA is committed to some task by a human agent, and whereby its failure exposes the HA to harm, trust was arguably betrayed rather than simply let down. Whether an AA then is trusted or simply relied upon may depend on the importance of the task or object with which it has been entrusted.

Finally, it is necessary to locate the moral value of trust, particularly within the framework of IE. Trust is popularly claimed to reduce complexity in interactions between agents, reducing suspicion and opening up possibilities for action not otherwise engaged, it facilitates co-operation and co-ordination under conditions of uncertainty, enhances relationships (intimate and professional), and over time it can produce cohesion in communities (Nissenbaum, 2001b, pp. 105–108; Taddeo, 2009). Trust facilitates the sharing of information too, and therefore plays an important role in generating knowledge (Taddeo, 2010a). Trust then can strengthen the cohesion of multi-agent systems by functioning as a kind of lubricant that supports co-ordination between the constituting agents—a multi-agent system where agents can delegate tasks or actions towards goals without fear or suspicion is one that can function more effectively, it is one that facilitates positive interactions and outcomes that may otherwise be unachievable. A multi-agent system where trust flourishes is one where the power of co-operative action and cohesion can be harnessed towards contributing to the flourishing of the infosphere as a whole. In this way it can reduce moral inertia. Trust then has a powerful instrumental value. Distrust (whilst a rational response to negative assessments of trustworthiness that can save individuals from betrayal) is quite the opposite, and can impede this cohesion and co-operation that trust supports. It can weaken moral resilience and render multi-agent systems ineffective, thereby increasing its vulnerability to entropy.[156]

[156] It might be noted though that trust can of course strengthen multi-agent systems that are formed with evil goals (for instance, crime syndicates), therefore it is a value that is only as beneficial as its overall operating environment and the relations it qualifies. On balance however, when situated as a second order property of relations geared towards morally neutral or positive

In the preceding, a descriptive account of trust was offered, that is, it is a second order property of a first order relation where A delegates an object C (or task) to trustworthy (the trust is justified based on relevant criteria for rational assessment to be made) B. Normatively, where B fails to execute its task (C), A is betrayed by B, justifying loss of trust between A and B.

7.3.2 Slándáil-type EMIS and Trust

7.3.2.1 Information and Disinformation

The first task here is to investigate the possible impact of information/disinformation on relations qualified by trust. To structure such an analysis, the framework of trust qualified relations as described by Grodzinsky and Wolf (2011, p.25) will serve usefully to examine how the occurrence of information/disinformation on social media in natural disaster response affects relations between agents qualified by trust under the categories of HHE-trust/HHP-trust, HAE-trust, and AHE-trust.

HHE-trust/HHP-trust

The first broad category of trust qualified relations to examine is that between human agents as mediated by a digital environment (social media), and also human agents in the physical environment. Firstly, in a disaster situation, emergency managers utilising social media will trust the disaster affected, and persons otherwise with relevant information to the situation, to assess the situation and post factual reports on social media that can be verified and acted upon. This degree of trust could be quite limited to begin with, as an emergency management participant in interviews pointed out, emergency managers may first dispatch their own "trusted asset" to verify and assess the reported situation. This does not indicate an absence of trust, as the asset itself is one of few resources and there may be an opportunity cost in dispatching him/her to the area of the reported incident—the emergency manager still delegates the responsibility of intelligence acquisition to the concerned citizen and may dispatch one of limited assets knowing that there remains a risk that there is no incident and the trusted asset could have been dispatched elsewhere. Each instance where the trusted asset can verify the reported information is likely to increase trust in the citizen reporter

---

ends, it is a value that stands to improve the efficiency MASes and the infosphere that they constitute.

as an abstract entity, however each instance where the citizen's report proves to be disinformation or misinformation is likely to undermine this trust.

Placing trust in the citizen reporter can be strained under the circumstances of social media where relative anonymity can be prevalent and social forces and cues are largely absent, as discussed—it can be difficult to find criteria to justify trust. However the presence of an AA such as Slándáil can to some extent justify trust in the information received and by extension the citizens who are generating it. A valid criterion to justify trust in this particular scenario would be location, a criterion, in the human context, which also gives way to inferences of characteristics, and shared values and norms—the reporting citizen will probably be based in the emergency manager's locality. Slándáil, and similar systems, filter information that is generated from specified locations, which gives grounds for the emergency manager to trust it, and the citizens who have reported it (who are likely from a context familiar to the emergency manager). The information then might be said to be trustworthy, and the citizen reporter trustworthy, as it has been *referred* by potentially trustworthy Slándáil.

Of course, the performance of the Slándáil system may also reinforce or undermine trust between the emergency manager and the developers/vendors of such a system—if it fails to deliver credible, actionable information and instead consistently delivers false information, the emergency manager's trust in the system (as will be explored soon) will be undermined, and their trust in the system's creators. As a result, the emergency manager may elect not to do business with the system's creators again.

If false information regularly dominates social media feeds during disasters and causes emergency managers to repeatedly dispatch assets to false alarms, this method for intelligence gathering, and by extension the citizen reporter, might be deemed untrustworthy. The risk of trusting the citizen reporter over social media (where, after all, digital wildfires can spread at alarming rates), might be too steep to justify trust, and it may be more rational and less wasteful to stop utilising social media as an information source during natural disaster. This persistent viral spread of false information potentially also has the capacity to sow seeds of distrust between citizens on social media, who will trust each other to relay accurate information about the event unfolding around them—if the citizen/disaster affected is repeatedly mislead by their peers then their trust in their peers will be undermined. They will have less credibility, and social media itself as a source of information will have less credibility.

The inverse of this is the citizen affected by the natural disaster, who must trust the emergency responder with their well-being. The citizen trusts the emergency manager with tasks related to ensuring their continued safety in the event of emergency. Should the emergency manager choose to take action based on false social media reports (for instance, divert resources to where they are not needed), and thereby waste time and resources that would be better deployed elsewhere (and thereby potentially rendering the citizen perhaps urgently in need of them vulnerable to danger), trust in the emergency manager and the agencies s/he represents will be undermined. Such a loss of trust can be quite dangerous where the citizen may view the emergency manager as so untrustworthy that they, for example, ignore evacuation orders.

Although when defining trust here it related to the delegation of tasks/objects towards specific goals, it is fair to assume that poor performance in a given activity will affect trust assessments relating to other goals—if a citizen perceives emergency management agencies as being untrustworthy in one instance of a trust qualified relation, this may impact their trust assessment with regards to their competence and trustworthiness in the execution of other goals that are a part of their roles, that is, their reputation could be undermined. Citizens that deem emergency managers (representing in many cases the police) ineffectual or incompetent, will strain to place trust in these statutory agents when they need them in other situations (for instance, in reporting crime to the police). In this regard, failure of HHE-trust may impact HHP-trust.

False information emerging from ignorant or malicious agents can undermine trust between emergency managers and citizens, between citizen and other citizens, and between citizens and emergency managers, and in these complex interrelations both HHP and HHE trust can be damaged. The MAS is weakened as trust diminishes in this way, the quality of interactions that makes it efficient overall may deteriorate, and as a result moral resilience may fall, exposing the MAS and the larger infosphere to the forces of entropy to which the MAS may be less able to effectively coordinate against. To avoid this, sources of distrust must be mitigated.

HAE-trust

Human agents form trust qualified relations with AAs to the extent that a Slándáil-type system is an AA that has been delegated the task of delivering relevant information during natural disasters to emergency managers.

It can be difficult to justify trust in an AA such as Slándáil, though the reputation of its creators, along with its predictability and reliability (established through testing), and transparency of how the system operates (Grodzinsky, Miller and Wolf, 2011), might serve as effective criteria.

The AA (such as Slándáil or similar) is trusted with collecting and presenting relevant disaster related information from social media to emergency managers, towards the end of assisting decision-making in disaster management. If it fails to produce significant relevant information, or presents inaccurate information, it does not necessarily live up to expectations of producing efficiencies by sorting signals from noise and presenting actionable information. If the AA presents irrelevant information (which is not a strict focus here, but broadly relevant to the topic of trust) that is simply noise, its benefits in terms of reduced manpower are somewhat neutralised and it has failed in its goal. If it presents false or inaccurate information from untrustworthy sources which the emergency manager might assume are trustworthy, the emergency manager may make a wasteful and dangerous decision on resource allocation—they would have been betrayed after making a decision based on the output of the AA. To this extent, the AA becomes somewhat accountable for agents that generate misinformation or disinformation. It is the expectation of the emergency manager that actionable information will be relayed—should the output of the system be irrelevant or inaccurate, the AA has failed (at least partially) at the task to which it has been entrusted. So, not only is trust in the citizen reporter implicated, but trust in the AA that acts as a mediator between the emergency manager and the citizen reporter.

Of course, some error might be expected, without any risk there is no basis for trust, however should the failures of the AA be substantial and consistent, justification of trust is not tenable. Simply put, if the AA does not perform to its specified purpose in a manner that effectively assists emergency managers make potentially life-saving decisions, even if its failure can be traced to misinformation or disinformation being generated by human agents, it is not trustworthy.

AHE-trust

Perhaps one of the more novel aspects of trust across digital environments is the capacity of AAs to both place trust in other AAs and HAs. The Slándáil system simulates trust in human agents to some limited degree, even if it does not model and utilise this

trust with the mathematical precision of which AAs are capable (Taddeo, 2010b; Turilli, Vaccaro and Taddeo, 2010; Grodzinsky, Miller and Wolf, 2011). It uses criteria to assess the relevance of messages posted on social media, and the relation assumed between AA and HA takes on a trust-like property when the AA deems that the criteria relating to the relevance of the user generated message is fulfilled—in determining the relevance of the message, the AA will analyse location (though as stipulated by the emergency manager), and textual or image content based on terminology databases and such. In determining relevance based on these criteria and allowing the message to be filtered through and presented to the emergency manager, it has placed some trust in the human agent, trust that the information generated by the human agent is true and accurate. This is, however, a limited demonstration of trust, as whilst the AA does utilise criteria that *de facto* establishes whether the information is relevant (a trust assessment of sorts), and *de facto* the social media user trustworthy, no precise risk assessment is conducted and the AA loses no trust in the HA in the event that it has been deceived. Negative outcomes associated with this rather superficial imitation of a trust qualified relation then will impact HHE-trust/HHP trust, as discussed earlier, and HAE-trust. The trust that the AA places in the HA at fault cannot really be said to be undermined if the AA cannot act on the betrayal, that is, if the AA cannot factor in this betrayal of trust in its future interactions with the HA at fault in determining whether or not to present the user's post to the emergency manager or roll it into aggregated analysis.

As a mediator between emergency manager and social media user (including the disaster affected), it is problematic that the AA's capacity to make trust assessments and act on them is constrained by the lack of functionality to do so, as is the case with the Slándáil system. The AA's limited capacity to fully utilise trusting operations (assess trustworthiness and essentially update a trust index in determining future engagements with untrustworthy social media users) is an impediment to the system's integrity, and credibility as an effective decision support system, and in potentially presenting misinformation or disinformation to emergency managers (particularly if this is a recurring problem), its own trustworthiness is undermined. The system is essentially, without refined trust assessment criteria, only as trustworthy as social media users.

Though participants (both technical and those involved in emergency management) in the interview stage of this research emphasised (and offered theoretical musings in response to questioning) the need for the system to be able to determine the credibility

of information isolated and presented by the system, at time of writing no technological method has been established or implemented in the Slándáil system. The system has been delegated a task that might normally be assigned to multiple HAs, who would execute it with good (ideally) judgement and by trusting based on appropriate assessments. As the system is delegated with a formerly human task, one which benefits from trust assessments and decision making, it too should be able to emulate or even improve upon this assessment task (enhanced by its computational capacity). By maximising the system's ability to fully implement and act on trust (through effective assessment and amended assessment based on betrayal or success of the trusted), it can potentially usefully mitigate the threat of misinformation and disinformation.

The first step in improving the system's implementation of trust is through refined and precise assessment criteria that enable it to automatically make an assessment of the credibility of the information it processes. Whilst such a refined assessment process is not exemplified at present in the Slándáil system, fortunately research has been conducted elsewhere that can be drawn upon here.

Automated systems can be used to assess the credibility of events discussed on social media based on certain features. These features are synonymous with trust criteria, they involve analysis of characteristics of the social media user and characteristics of their messages, so a trust assessment of the social media user is conducted by the AA in order to gauge the trustworthiness of the social media user and infer message credibility based on this. In their research, Castillo, Mendoza, and Poblete (2011 p. 682) identify features that can effectively contribute towards event credibility assessment including features from their top-element subset,[157] and propagation (retweets for instance) subsets. The researchers emphasise that the presence of URLs in tweets, and "deep propagation trees" can indicate the credibility of a reported event. In addition, they state that "[a]mong several other features, credible news are propagated through authors that have previously written a large number of messages, originate at a single or a few users in the network, and have many reposts," (Castillo, Mendoza and Poblete, 2011, p. 682).

[157] The researchers' top-element subset is described as "consider[ing] characteristics of the text of the messages. This includes the average length of the tweets, the sentiment based features, the features related to URLs, and those related to counting elements such as hashtags, user mentions, etc." (Castillo, Mendoza and Poblete, 2011, p. 682).

Other researchers have explored event credibility in other contexts (whilst the former researchers investigated rumour detection on Twitter, Yang, Liu, and Yang investigated it on Sina Weibo) (Yang *et al.*, 2012). In Yang, Liu, and Yang's (2012) research, the researchers assess event credibility using content-based, client-based,[158] account-based (that is, personal characteristics of the user), propagation-based, and location-based features. Yang, Liu and Yang (2012, pp.6-7) found that account-based, and content-based features were effective in credibility assessment.[159] The researchers also report the client type and user location (proximity to event) are effective in classification of event credibility (Yang *et al.*, 2012).

Other research has added to the preceding by attempting to determine the criteria, or features, that can be used to assess the credibility of images circulating social media as they pertain to events. Gupta *et al*. (2013) use a similar methodology as the preceding cases to determine whether automatic classifiers can correctly identify false information (in this case, tweets containing image URLs of misleading or fake images). The researchers assessed features including "[s]ource or user level features" (number of friends, followers, and status messages) and "[c]ontent or tweet level features... (e.g. words, URLs, hashtags) and meta-data (e.g. is tweet reply or a tweet) related to it"(Gupta *et al.*, 2013, p. 5). In this case, Gupta *et al.* (2013, p. 6) had more success using tweet content to identify false image URLs than with user level features.

This research indicates that numerous features, used as proxies for criteria in trustworthiness assessment, can be utilised by automatic classifiers to determine event credibility. Essentially, on a superficial level, the classifiers described implement a trust assessment of social media messages and users based on these criteria in order to establish whether information is credible. Such classification can be used either to make multiple trust assessments across a range of messages in aggregate to determine the

[158] The application the social media service was used on, for instance a mobile or desktop client.
[159] Relating to account-based features in particular, Yang et al (2012, pp. 6-7) note:

> Most of the account-based features are user's attributes, so it is effective to detect the false rumours by microblog-account's features, like weather the user's account is verified, the number of friends, the time span between its registering time and the posting time. For instance, if one who is verified by Sina Weibo and has a large number of friends (fans), then the microblogs posted by this account are rumours with a small probability. Contrary to this scenario, if one is just registering with little friends (fans), default or fake avatar, and not verified by the official service, then the message posted by this account is false rumour with high probability if this microblog related to a controversial event.

likelihood of a reported event being true, or on a case-by-case basis to establish the credibility of individual messages and social media users. Such a methodology (or if an independent component of an overall system, an AA), applied in the context of a Slándáil-type system, could plausibly be used to exclude messages from sources deemed untrustworthy from presentation to emergency manager.

The implementation of trust in such a scenario is not complete unless the AA can act upon betrayal and base future interactions with social media users on this betrayal. It might for instance attach a trustworthiness score to users that increases or decreases based on successful interactions. This may require the input of emergency managers or developers, who may need to flag messages presented as credible by the system as false, enabling the system to conduct an updated risk assessment of particular social media users during future events. A full implementation of trust may need to go beyond refined criteria assessment, and trust may need to be modelled with explicit mathematical weighting and AA reflection as proposed by Taddeo (2010b).

This is a task for developers and programmers going forward, however it is a worthwhile one. The AA that can successfully implement trust, one which can make trustworthiness assessments and risk calculations of social media users and the information they generate, is one that can be trusted and can be used (in identifying information and disinformation) to mitigate against the viral spread of rumour, something which has the potential to cause harm. The trusting AA can help preserve the integrity and cohesion of the MAS and improve moral resilience in the environment. The converse of this, an AA that fails to make effective trust assessments, undermines trust in the MAS and weakens the integrity of the MAS, and loses credibility as a useful decision support system in the process.

7.3.3 The Problem of Function Creep

Systems that analyse data are highly adaptable, and are ripe for application in contexts outside of those for which they were originally designed.

Function creep is not in itself ethically wrong, and might even be morally good where it achieves positive ends. Any substantial deviation from a system's initial functionality, however, should be subject to additional ethical analysis.

In the case of Slándáil-type systems, to the extent that they are designed to monitor social media exclusively for messages that assist decision-making in natural disaster

management, function creep can occur quickly and may not cause distrust by citizens (assuming that there is trust to begin with). An expansion of terminology databases towards the goal of monitoring a greater range of natural disaster events for instance would be an example of function creep (to the extent that it represented widening functionality), and should be welcomed. Even if the system were later designed to include technological disasters, given the minimal qualitative difference between a natural and technological or man-made disaster, this may not necessarily be an ethically undesirable instance of function creep, and may not necessarily affect the system's impact on trust-qualified relations.[160] After all, natural and technological disasters can intermingle.[161]

Arguably the greatest danger arising from function creep of such systems is where their functionality and purpose mutates from relief and response towards crime prevention or countering civil disturbances. Compelling reasons can be offered for widening the scope of functionality and purpose of Slándáil-type systems to include not just environmental threats but human threats too, and within the context of natural disaster. After all, although the threat of looting, for instance, is said to be overstated (and sometimes will simply entail appropriation of necessities as opposed to the opportunistic theft of luxuries), it remains a risk in times of dearth, and a threat to public order, private property and safety (Auf Der Heide, 2004; Trainor, Barsky and Torres, 2006; Quarantelli, 2008). Indeed, in the aftermath of disaster the very *status quo* can be threatened by human agents as the risk of civil conflict is exacerbated (Nel and Righarts, 2008).

Function creep, where the functionality of Slándáil-type systems extends to the monitoring of human behaviours towards public order or crime prevention purposes is plausible, and an both interview participants representing emergency managers expressed an interest in its use outside of the natural disaster context. While one participant acknowledged that this was something that required some ethical contemplation, the participant did assert that additional functionality would indeed be useful. The ease with which the Slándáil system (the Social Media Monitor in particular)

[160] This to the extent that the system is utilised towards relief and response, and not towards the identification of suspects (for instance in terrorist incidents), which would require additional ethical analysis, and in practice, turning the purpose of the system towards recrimination, might have far reaching impacts on trust-qualified relations.

[161] Consider the 2011 Japanese Earthquake and Tsunami and consequent nuclear emergency in Fukushima.

could have functionality added was confirmed by a technology participant, who explained that the dictionaries could be updated with relative ease, as a goal of the development of the system was for it to be customisable to meet the needs of emergency managers. An expansion of dictionary terms would enable the SMM to collect social media messages outside of natural disaster scenarios to the extent that included terms were relevant to other events or situations.

Such a creep needs to be considered with care, as it may have implications for the autonomy and trust of human agents across the physical and digital regions of the infosphere. The first thing to consider is that it would appear to provide a fertile ground for lateral surveillance, where human agents' own biases and prejudices could manifest into active suspicion of "the other", "strangers" or even simply people they dislike. It is plausible that in this scenario social media could become a medium for posting aspersions or images about persons deemed suspicious, and thereby facilitate invasions of privacy and mischaracterisations of innocent individuals. Trust would not be only be undermined as much as encouraged to never flourish, between human agents, and in an environment of such uncertainty, persons who might suspect that they are under the scrutiny of others may not be able to trust others with the negative object of not violating their privacy. As highlighted in the work of Marder *et al.* (2016), the simple existence of social media is enough to cause a chilling effect that makes persons moderate their real-world behaviour, and as argued above, this exposes a certain existing distrust between persons and their peers—human agents' autonomy is challenged by the perceived possibility of some social sanction arising from undesirable behaviour, people simply do not trust their peers not to judge them. If a widely used, transparent criminal reporting system is embraced, such an effect would arguably be escalated.

Conversely, a transparent social reporting system, a sort of digital community watch that facilitates open lateral surveillance could have boons and increase community cohesion and trust (suppose the persons might trust neighbours to actively protect their safety, rather to not violate their privacy). Purenne and Palierse (2017) after all observed pros and cons to the phenomenon of community based surveillance. Just how trust in society is impacted by such function creep into the realm of surveillance to combat crime or public disorder will likely depend on pre-existing factors, such as pre-existing community cohesion, and the integrity of statutory agencies and the Government.

The possibilities for surveillance of public social media feeds are extensive, and function creep could of course extend into non-emergency scenarios. Given the potential for end-users to customise the dictionary, and therefore the filtering parameters of the system, a system such as Slándáil could plausibly be used to track crimes like internet harassment. This insertion of a statutory policing authority into the digital space could potentially contribute to chilling too, as persons may balk at sharing information if they believe this will find them in trouble with the law, where they cannot trust law enforcement not to sanction them. This aspect of chilling behaviour was demonstrated in the research of Penney (2016). In this scenario however, the chilling could also be a deterrent from doing harm where there is increased certainty that the commission of a crime will be detected and punished. In such cases, the public may trust the statutory agency (and the AA) to protect their safety, and perhaps justifiably so. Again, pre-existing factors may influence the impacts on trust. Pre-existing trust in police and statutory authorities to conduct their mission with integrity will likely influence the public trust in the statutory agency as mediated in the digital space.

Function creep may also be synonymous with overtly unethical system use or practices. For instance, if a Slándáil-type system is used in a manner that promotes the inappropriate flow of personal information (let us assume for the sake of argument that statutory agencies or the system creators commercialise unredacted collected tweets to a private marketing enterprise), or if it uses social sorting or profiling to target and persecute particular ethnic minorities. Overtly unethical use can impact public trust in the system and end-users (or creators), citizens will make a negative assessment of the implicated agents' trustworthiness in protecting their privacy or safety and this could also lead to chilling (perhaps even civil conflict at an extreme as far as social sorting goes).

Outcomes under this category are especially difficult to predict, and a vast array of variables are involved—consider pre-existing trust in the authorities or private corporations by citizens; the pre-existing trust may differ substantially between autocratic and democratic regimes for instance (and justifiably so). The area of trust and function creep is one which will especially require more research, particular under the discipline of social science, after the deployment of the system in order to quantitatively and/or qualitatively test its impact on trust in a range of conditions. The present

research can only hope to anticipate potential arising risks, and indicate solutions to mitigate the potential of these harmful risks.

7.4 Fiduciary Theory and Trust

7.4.1 Trust and the Fiduciary Relationship

This section focusing on human rights represents a departure from the approach taken in previous chapters. Unlike privacy, or a discrimination analysis based on the concept of justice, there is no human right to trust. Trust is emphatically not a human right, however it plays an essential role in the mechanics of human rights to the extent that they form the blueprints of a regime of secure and equal freedom as established and maintained by the state in the fiduciary-subject relationship. Once the position of trust is understood in this relationship, it becomes apparent that it facilitates quite a broad human rights analysis in this latter portion of the current chapter.

As explained to some extent in Chapter 2, trust is a useful explanatory concept that underpins the fiduciary-legal subject relationship. Under Fox-Decent's (2011) conception of trust as it operates in Fiduciary Theory, it is quite different from described in the preceding section, it is more minimal. Under Fiduciary Theory the fiduciary is entrusted with its legal powers by the law (or fiduciary principle) and not directly by the legal subject (Fox-Decent, 2011). It is a slight modification of the basic definition of trust offered by Baier (1986) where A trusts B with object C, to "...the law entrusts an actor, B, to do C on behalf of A," (Fox-Decent, 2011, p. 106), therefore the law exists as a fourth element in the traditionally three part relationship.[162]

There is a significant variance in this use of the concept of trust as explored this chapter, and it runs counterintuitive to an interpretation of trust as something that is indivisible from agency. The fiduciary principle itself cannot qualify as an agent, it is at its core monolithic, even if the content of state duties in establishing and maintaining a regime of secure and equal freedom is subject to change (considering that human rights form the blueprints of this regime, and are subject to revaluation and the emergence of new

[162] According to Fox-Decent (2011, p. 106):

> ...the fiduciary principle entrusts the state to establish legal order on behalf of the people. The state in turn exercises power on the basis of the people's trust (the public trust) precisely because the fiduciary principle entrusted the state with public powers on their behalf. Thus, trust plays a central role in the state-subject fiduciary relationship even if the subject rejects the state's claim of authority over her and distrusts the state.

rights). This need not undermine the theory, if "entrust" is read in a more practical and literal (if admittedly threadbare) sense than philosophical, that is, it is the "[a]ssignment [of] the responsibility for doing something to (someone)" (Oxford Dictionaries, no date). The principle does not *trust* the state with power, it *entrusts* it.

The fiduciary principle entrusts and authorises the state to use its power for the public benefit in this direct and literal sense. The law places legal powers into the hands of the fiduciary and gives it legal authority, within the parameters of the fiduciary principle, to exercise powers to the sole benefit of the beneficiaries, legal subjects, who are incapable of doing so themselves and are vulnerable to the administrative power the fiduciary wields. The state exercises its power on the *basis of the people's trust*, though this is trust in the abstract, it is a *presumptive trust* that the state must act on even if its subjects do not in fact believe that it is trustworthy (Fox-Decent, 2011a, pp. 107–109).

Trust does bear into the fiduciary relationship in a more traditional sense, the legal subject is provided with justification to trust the fiduciary state (whether or not the reality is that s/he believes the fiduciary trustworthy), as Fox-Decent (2011, p. 106) argues:

> From a legal point of view, to say that a beneficiary can trust and rely on a fiduciary means that the law requires the fiduciary to exercise power on the basis of the beneficiary's trust. Because the fiduciary is legally required to act in conformity with obligations that flow from the trust-like nature of the relationship, the beneficiary is assured that the law protects his entrusted interests. Thus, the beneficiary has a legal basis to trust and rely on the fiduciary, whether or not he in fact trusts her... In sum, the fiduciary obligations assumed with every exercise of fiduciary power give the beneficiary reason to trust and rely on the fiduciary, since those obligations make the fiduciary liable to the beneficiary should the fiduciary breach them.

Fox-Decent (2011, p. 107) further appeals to the notion of "automatic and unconscious" trust that defines relationships of asymmetric power, such as between parent and child, where trust is not necessarily explicitly expressed between parties but is implied. This parallels with and supports the Kantian explanation of the fiduciary relationship, whereby "...persons have an innate right of humanity that can place others under obligation without any act being required of the right-holder" (Fox-Decent, 2011, p. 107). This approach to and usage of trust need not be viewed as being contrary to the explanation provided previously here, but rather as an explanation of the manifestation of trust under alternative conditions, and under different parameters, not directly comparable to the exploration of trust at that point.

Under Fiduciary Theory (and on the basis of trust) the state is expected to provide a regime of secure and equal freedom under the rule of law, and to announce and enforce the law—this creates a reciprocal relationship with legal subjects, who are (both morally and legally) required to obey the law to the extent that it is indeed law and not contrary to the features and requirements of the fiduciary duty (the simple decision or rule that does not meet the quality of law is not truly law at all) (Fox-Decent, 2011). Where the state fails to comply with the rule of law, and acts contrary to the fiduciary duty, its legitimacy is challenged (Fox-Decent, 2011). Where the state fails to comply with the rule of law, and where its actions are issued as simple rules or decisions (not law), it risks violating human rights, which are among its fiduciary obligations. Illegitimate rules or decisions do not command obedience, and may even warrant resistance (Fox-Decent, 2011). The state that defies its fiduciary duties through arbitrary power, and neglects the provision of a regime of secure and equal freedom under the rule of law (through a blueprint of human rights), violates the public trust under which it holds its power (Fox-Decent, 2011). This state betrays the public trust, and may, given sufficient circumstances, demand accountability from a public that may not even be obliged to obey the "laws" that it issues.

So the state that violates human rights betrays the public trust. The state that betrays the public trust undermines its legitimacy, and invites resistance. The betrayal of this trust has real consequences.[163]

If human rights violations then represent betrayals of the public trust, the remainder of this section can be used for very broad human rights analysis. Already covered were the subjects of privacy and discrimination, therefore these need not be revisited explicitly. The foregoing analysis demonstrated that there is a risk of chilling effects from expanded surveillance, and such chilling, as noted by Scheinin (2009), has implications for rights including the freedom of expression, association, and assembly. The remainder of this chapter will advance the human rights analysis to these three rights, on the basis that they are potentially at risk in the deployment of a system such as Slándáil, and their violation would represent a breach of the public trust.

[163] Research has shown that the aggrieved public will challenge states where human rights are violated, and this trust broken even violently (Thoms and Ron, 2007). International organisations, as the secondary guarantor of human rights are also likely to intervene where this trust has been broken, on behalf of the victimised public that has been victimised, as a long history of sanctions levied against rogue regimes evidences.

7.4.2 Freedom of Expression

Freedom of expression holds intuitive importance in an account of human rights under Fiduciary Theory. Freedom of expression promotes democratic culture, it fosters and protects public political (and cultural, even moral) debate, and artistic expression that in turn can subvert and challenge the state where it is failing in its fiduciary duty (Balkin, 2004; Harris *et al.*, 2009, p. 443). The freedom to impart and receive information is valuable in a democratic (or non-democratic especially) society as it can contribute towards cultures of transparency. It allows journalists to cast light on perhaps otherwise opaque practices of the state or other powerful actors within the state, and allows citizens to make informed decisions about things that affect their lives.[164] Where freedom of expression is not unreasonably restricted, citizens are in a position to push back against forces of domination and instrumentalisation (corrupt governments or corporations as examples), which is intuitively a necessary freedom to help protect the regime of secure and equal freedom under conditions of non-instrumentalisation and non-domination promised by the fiduciary relationship with the State.

In international law, freedom of expression is protected by Article 19 of the *ICCPR* (1966). In the *ECHR* (1950), the right is enshrined Article 10.

In the case law of the ECtHR, broad protection is granted to freedom of expression, including the transmission and receipt of information regarded as unpalatable or offensive, and "[t]he scope of protection under Article 10 is to be broadly interpreted so as to encompass not only the substance of information and ideas, but also a diverse variety of forms and means in which they are manifested, transmitted, and received" (Harris *et al.*, 2009, pp. 444–445). Numerous forms of expression are granted protection by Article 10, including political, civil, commercial and artistic expression (Harris *et al.*, 2009).

The ECtHR has also established that states hold a positive obligation to protect freedom of expression (though not without qualification), including ensuring some degree of freedom of expression in relations between private persons (Harris *et al.*, 2009, p. 46).[165]

[164] An informed public for instance is in a better position to grant political authority wisely to governments during elections, or challenge a corrupt government that occupies the role of fiduciary.

[165] Citing *Ozgur Gundem v Turkey* [2000], and *Appleby and Others v UK* [2003], Harris, O'Boyle, and Warbrick (2009, p.46) outline the Court's approach to positive obligations in this regard:

Ozgur Gundem v Turkey [2006] is an illustrative example of where the Court found a need for a positive obligation to protect freedom of speech vis-à-vis relations between private actors. Here, a Turkish newspaper was subject to violence and intimidation which led the murders, assaults, and arson (Harris *et al.*, 2009, p. 447 citing *Ozgur Gundem v Turkey,* [2006]). The Court found that Turkey was responsible for investigating and protecting the journalists in this situation, and its failure to do so was tantamount to an Article 10 violation (Harris *et al.*, 2009, p. 447 (Harris *et al.*, 2009, p. 447 citing *Ozgur Gundem v Turkey,* [2006]).

In contrast, in *Appleby and Others v The United Kingdom* [2003], the Court rejected the notion of a positive obligation for a "freedom of forum", finding that campaigners had no right to conduct their campaigning on private property (a shopping mall)—vitally, these campaigners had alternative options, or locations, to conduct this business (Harris *et al.*, 2009, p. 447 citing *Appleby and Others v The United Kingdom,* [2003]).

Freedom of expression is not unlimited, as the limitation clauses present in paragraph 3 of the *ICCPR* (1966) and paragraph 2 of the *ECHR* (1950) show. As with privacy, a number of grounds serve as sufficient justification for limitations to the freedom of expression, though as with privacy, such limitations are subject to being prescribed by law, pursuing a legitimate aim, and being necessary in a democratic society. Some forms of expression benefit from no protection at all and in fact the state has a positive obligation to combat them, such as hate speech as demonstrated in *Gunduz v. Turkey* [2004] (Council of Europe Research Division, 2015, pp. 19, 54). Where the freedom of expression is in tension with other rights, compelling public interest must be offered in favour of freedom of expression, as in the cases of *Axel Springer AG* [2012]*, and Von Hannover v. Germany (No. 2)* [2012] for instance, the Court considered the following criteria in balancing freedom of expression against privacy, "...contribution to a debate of general interest, whether the person concerned is a public figure, the subject of the report, the form and repercussions of the publication and the severity of the penalty imposed"

> In determining whether a positive obligation to act exists in a particular situation 'regard must be had to the fair balance that has to be struck between the general interest of the community and the interests of the individual. The ambit of the state's positive obligation varies, depending on considerations of distributive justice and the equitable allocation of resources required for different administrative tasks. Relevant factors are: the kind of the expression rights at stake; their public interest nature; their capacity to contribute to political debates; the nature and scope of restrictions on expression rights; the availability of alternative venues for expression; and the weight of countervailing rights of others or the public.

(Council of Europe Research Division, 2015, p. 21 citing *Axel Springer AG* [2012]*, and Von Hannover v. Germany (No. 2)* [2012]).

The state, in implementing limitations to the freedom of expression, must be careful not to unreasonably stifle that very freedom. Severe restrictions on the freedom of expression or penalties arising from the imparting of information were found by the Court to risk a chilling effect incompatible with freedom of expression, and as expressed in *Mouvement raelien Suisse v. Switzerland* [2013], the state is required "...to choose the means that cause the least prejudice to the rights in question" (Council of Europe Research Division, 2015, pp. 33–34 citing *Mouvement raelien Suisse v. Switzerland,* [2013]). Prison sentences, for example are a particularly extreme example of measures that may contribute to a chilling effect, as found in *Belpietro v. Italy* [2013], and even smaller sanctions such as fines may suffice to cause risk of chilling, as was argued in the case of *Morice v. France* [2015], and *Eon v. France* [2013] (Council of Europe Research Division, 2015, p. 34 citing *Belpietro v. Italy* [2013], *Morice v. France* [2015] and *Eon v. France* [2013]).

The case law and finer points of freedom of expression are extensive and this has offered but a superficial examination, nonetheless enough has been said to allow for a sufficient analysis of the issues in what follows.

7.4.3 Freedom of Association and Assembly

Freedom of association "...involves the freedom of individuals to come together for the protection of their interests by forming a collective entity which represents them" (Harris *et al.*, 2009, p. 525) and assembly "...the right of individuals to assemble and to associate for the furtherance of their personal interests, be they economic, social or cultural" (Harris *et al.*, 2009, p. 515). Whilst these two rights are distinct, "...they share the objective of allowing individuals to come together for the expression and protection of their common interests" (Harris *et al.*, 2009, p. 515).

The right to mobilise with others and form a stronger entity than any one person can be quite essential in enabling the public to resist domination, to collectively protect their rights,[166] and to influence states on the direction of their policy.

[166] Consider trade unions and how they can essentially help their members resist instrumentalisaton in the labour context.

In IHRL, the right to freedom of association is enshrined in Article 22 of the *ICCPR* (1966). Freedom of assembly is enshrined separately in Article 21 of the same document (*ICCPR*, 1966). Both freedoms share Article 11 of the *ECHR* (1950).

In the case law of the ECtHR, the meaning of association is autonomous and whether or not an association has legal recognition in a given state has little bearing as to whether or not it will be recognised by the Court, as "...the fact that a substantive co-ordination of activities of individuals is not recognised in the national 'association' will not necessarily mean that freedom of association is not at stake under Article 11" (Harris *et al.*, 2009, p. 526). Salient examples of associations include political parties and trade unions.

Freedom of association imposes positive obligations on the state, therefore the conditions for the effective operation of such associations must be present (Harris *et al.*, 2009, p. 536). As with freedom of expression, this also requires states to protect associations from violence and intimidation, as demonstrated in *Ouranio Toxo v. Greece* [2006] (Harris *et al.*, 2009, p. 536).

Harris et al. (2009, p. 516) argue that "Article 11 protects the right to freedom of peaceful assembly as a 'fundamental right', whether it is exercised for political, religious, or spiritual, cultural, social, or other purposes. It covers private and public meetings, including marches, demonstrations, and sit-ins." Such a freedom is powerful, allowing persons acting collectively to gain the attention of the media where they do not benefit from the power and influence of established parties (Harris *et al.*, 2009, p. 516). The right is considered in connection with Article 10, freedom of expression, to the extent that assemblies are allowed to express and pursue their goals within reasonable confines without being censored or impeded by the state unnecessarily, insofar as their pursuits and expression is not contrary to the protection of other rights (assemblies for instance, cannot be expected to be permitted to incite hatred or violence) (Harris *et al.*, 2009, p. 516). It should also be apparent that both the freedom of expression and assembly are valuable resources for association, allowing associations to publicly gather and express their interests and goals in an open forum. Again, the state has a positive obligation to protect peaceful assemblies from violence, and, as with freedom of expression, there is no entitlement to a forum or venue insofar as private actors control possible venues for assembly (Harris *et al.*, 2009, pp. 517–518).

Both the freedom of association and assembly are, as with all rights examined in this research, subject to limitations outlined in their respective limitations clauses in the *ICCPR* (1966) and *ECHR* (1950), and subject to restrictions being prescribed by law, pursuing a legitimate aim, and being necessary in a democratic society. Associations can of course be a dangerous force under certain circumstances, where they seek violent ends or aims and incite hatred. However, interferences with the freedom of association are not taken lightly by the ECtHR.[167] Interferences in this right may be particularly tolerable where the means and methods of an association are contrary to democratic principles (Harris *et al.*, 2009, p. 534).

Public assembly, such as demonstrations, pose challenges for state authorities that are charged with preserving public order, which can be disruptive and can escalate into violence (Harris *et al.*, 2009, p. 516). As such, given the risks and prospects of disruption, the right may be restricted and regulated. Where there are requirements for notification or authorisation for assembly, the Court has not regarded such conditions as interferences in themselves, although refusals can be deemed interferences (Harris *et al.*, 2009, p. 520). The Court has upheld decisions in favour of applicants where the state refused permission for assembly, such as in the case of *Baczkowski and Others v. Poland* [2007], where a march and stationary assemblies raising awareness against the discrimination against minorities, disabled and women, was refused permission but continued nonetheless (Harris *et al.*, 2009, p. 520). As outlined by Harris *et al.* (2009, p.520 citing *Baczkowski and Others v. Poland,* [2007]), the Court:

> ... stated that a 'resumption of legality' of an assembly constitutes 'a vital aspect of effective and unhindered exercise of the freedom of expression'. Holding an assembly with an official ban in force held its risks and, in particular, there was no guarantee of official protection.

Harris *et al.* (2009, p. 520) note that the potential of refusals of authorisation can lead to a chilling effect as regards freedom of assembly which can affect participation.

Such a principle of tending to avoid chilling effects was evident in the Court's judgement in *Ezelin v. France* [1991], where a march deteriorated into violence and the complainant remained and refused to co-operate with police during subsequent questioning (Harris *et al.*, 2009, p. 524). Ezelin was reprimanded by a Court of Appeal to

[167] As argued by Harris *et al.* (2009, p.534), "[t]he refusal to register an association is a 'radical measure' preventing as it does, the association from commencing any activity. Likewise, the immediate and permanent dissolution of a political party is a 'drastic' measure and will be justified only in the most exceptional circumstances."

his capacity as a lawyer for not disassociating himself from the march or co-operating with the police (Harris *et al.*, 2009, p. 524 citing *Ezelin v. France,* [1991]). The Court found in the complainant's favour, finding, as noted by Harris *et al.* (2009, p.524 citing *Ezelin v. France,* [1991]) that "[a] 'just balance' must not discourage persons from making their beliefs peacefully known."

With a review of the basics of the freedoms of expression, association, and assembly complete, it is now appropriate to move on the analysis of these rights in relation to possible uses of Slándáil-type systems.

7.4.4 Slándáil-type EMIS and Freedoms of Expression, Association, and Assembly

7.4.4.1 Freedom of Expression

The Slándáil system itself, as conceived and intended for deployment, should theoretically minimally impact freedom of expression. The system is designed solely to collect and analyse information pertaining to natural disasters in order to provide decision support for statutory agencies responding to said disasters. Failing prosecution of individuals responsible for rumours that might emerge, the state in such a scenario provides no disincentive from expression, in fact, it can be argued to be encouraging it (through acting on information provided by citizens in a positive manner). Following the initial design and intent of the Slándáil system, the citizen is in fact encouraged to share information about events occurring within their environment, and there is little risk of sanction arising from this sharing insofar as the information is accurate and pertains to the disaster at hand.

Where the state elects to actively prosecute originators of false and misleading information that threatens to undermine the decision-making of emergency managers, and risks the misallocation of potentially life-saving resources, freedom of expression is implicated. The prosecution of individuals for imparting information on social media runs the risk of causing a chilling effect. Others may self-censor information or be deterred from sharing information if they perceive that authorities will pursue them for it. Nonetheless, the prosecution of individuals for false reports, on balance, need not represent an unjustified interference with freedom of expression given the potentially grave consequences involved.[168] Prosecution for the reporting of false information is on balance proportionate, and justifiable where prescribed by law. The necessity of

[168] Emergency resources deployed to where they are not needed can lead to lost lives if they are diverted from where they are urgently needed.

deterring such behaviour should be clear. Where the law benefits from forseeability and publicity, the risk of chilling should be mitigated and in practice should only deter the spread of false information rather than total self-censorship of potentially life-saving information.

In this case, the system is not directly implicated in the interference of the right as the direct measure in question that might represent an interference with the freedom of expression would be the legal action and prosecution of a false reporter/rumour monger. The system would be very present in the process as a whole however, acting as the mediator that brings the false information to the attention of the emergency manager (particularly where it contains the functionality to flag possible rumours, thereby enhancing statutory authorities' ability to identify and prosecute rumour mongers). To the extent that the system might actually help identify malicious actors that threaten the rights of their fellow citizens, the system can be used to help state authorities execute their fiduciary duties. In such a situation it would be important that the state authorities prosecute originators of information, and not simply those sharing rumours whose only folly is gullibility—blanket prosecutions would be disproportionate and would be an authoritarian response placing the public in a position of uncertainty, subjecting them to the domination of the state.

If the system were to be applied to situations outside of disaster response and expanded to include other objectives, such as public order, or the prosecution of crimes, freedom of expression may be more heavily implicated. As explained, the Social Media Monitor of the system is quite customisable, facilitating expansion of dictionary terms, rendering it technically open to the possibility of function creep. Where this creep occurs, people may be more traditionally the target of surveillance rather than simply as a source of information about the condition of the environment around them. Citizens can be used as either intentional or unintentional informants on crimes taking place around them, or even surveillance targets as crime suspects themselves.

The acceptability of this is contingent on context. The expansion of dictionaries to include different scenarios, or crimes, may greatly enhance the police's ability to detect crimes taking place both online and offline.[169] At least superficially then, the system

[169] For instance, the system could detect a geo-tagged image of a car that is being stolen, or it could be used the identify instances of harassment or racial abuse on Twitter for example.

could be used to help state authorities to execute their fiduciary duties across a broad range of contexts.

The use of the system, however, has implications for privacy, as discussed in Chapter 5 at length. Privacy should be viewed as having something of a symbiotic relationship with the freedom of expression, and the knowledge that one's privacy is under persistent interference would chill their acts of expression. If state authorities were to run a Slándáil-type system on a 24/7 basis in order to detect a spectrum of possible crimes, privacy would be under egregious interference, which may chill engagement with and posting on social media (each tweet or post might be deemed an act of expression). It is difficult to justify such broad use of these systems, as legitimate avenues currently exist for reporting the crimes that they can detect (one can report their stolen car or racial abuse in person or by phone), their use would be a disproportionate inference with the right to privacy, resulting in a disproportionate interference with the right to freedom of expression.

On the interference with the right to freedom of expression, it might be argued that alternative forums exist for expression free from the gaze of authorities (perhaps alternative websites that do not have open API access that Slándáil-type systems require). This argument is significant, even if a state that adopted it would be displaying hostility towards private actors.[170] That a state would undermine the interactions of what are very large communities (Twitter and Facebook, as described in Chapter 1, have very large memberships) because alternatives exist however is ultimately not acceptable—these websites are large community spaces with pre-existing relations between users, and pre-established norms that might not necessarily be transferable to alternative services, and much less so in real-life. Telling social media users that their freedom of expression is not being unreasonably interfered with because they can migrate to alternative services to express themselves is almost as unreasonable as telling townspeople who face impediments to the realisation of their free expression that any such impediment is not an unreasonable interference as they can simply move to a nearby village where this impediment does not exist.

Whilst it has been argued largely that function creep is impermissible, there are circumstances where it can be justified, particularly if the creep is within the context of

[170] For example, it would actively be undermining Twitter's business interests by essentially telling its users "go elsewhere or accept that we are watching you at all times".

natural disasters and other emergencies. Whilst in normal times the persistent monitoring of the social media space is unnecessary and disproportionate, dire circumstances may necessitate more restrictive measures in order to secure a regime of secure and equal freedom under the rule of law for all. Firstly, the expansion of a dictionary terminology to enable a Slándáil-type system to collect information with regard to crime occurring in real life in the immediate aftermath of a natural disaster is a reasonable step and one that should not interfere with the right to privacy, or expression, more than its standard use. Under circumstances where traditional reporting media (such as landline phones) may be unresponsive due to damaged infrastructure, or overwhelmed emergency call centres, this expanded functionality arguably meets criteria of necessity and proportionality in maintaining public order. As with privacy explicitly as argued in Chapter 5, derogation may be necessary in such contexts where the quality (or existence) of law is questionable.

There may be recourse for using expanded terminology dictionaries in other situations where traditional means of reporting are strained, even outside of natural disaster. The criteria that it be an emergency however is important, as such measures may stifle expression if implemented over longer time horizons. If the system can be used, and be demonstrated as being necessary (alternatives should be exhausted) to tackle civil disturbances and serious crimes that might follow an emergency (natural or man-made, perhaps even terroristic) it might be that the system with expanded functionality would be an essential tool in aiding the authorities' maintenance of a regime of secure and equal freedom.

7.4.4.2 Freedom of Association and Assembly

Slándáil as conceived and in its current form should have no implications for the freedom of association or assembly—it has been developed and trained for use using natural disaster terminology and can in no way be used in its current form, without modification, by authorities to interfere with these rights. In that case, a necessary pre-condition for it to have implications for these rights would be function creep.

Human rights professionals have indicated the risks of surveillance for the realisation of freedom of association and assembly.[171]

[171] Scheinin (2009, p. 14) for instance, argues that:

A state with a vested interested in quashing dissident voices in order to maintain its absolute authority may find a powerful tool in systems such as Slándáil when sufficiently modified to be able to isolate terminology occurring on social media associated with the activities of opposing voices, who may either represent an association or potentially be a collection of individuals in the nascent stages of forming some association, or in the process of planning an activity of peaceful assembly. It is not inconceivable that a state with poor respect for human rights would use a Slándáil-type system to bolster their surveillance of social media in order to identify and challenge internal (perhaps even external) threats to their hegemony—in the case of Slándáil (that is, the Social Media Monitor in particular), modifying it to sufficiently allow such surveillance may not be a complex task with regards to its customisability. Such suppression of activity with such importance in the democratic process, that allows persons to challenge the status quo or elements of the status quo, in order to foster dialogue and influence change, is incompatible with the fiduciary duty. Such rights could be harshly chilled where actively combated by the state.

The attempted suppression of associations and assembly need not be practiced exclusively by authoritarian or otherwise malicious regimes. Associations and groups engaging in acts of assembly may not represent valid interests, they may promote violent and subversive methods or be founded on the basis of discrimination against minority groups. The fiduciary state, following its duty to provide a regime of secure and equal freedom for all, would be required to combat the emergence of such groups or activities that are contrary to the rights of others, or otherwise exist to challenge and undermine the legitimate sovereign authority. A system such as Slándáil could allow the fiduciary state to monitor social media to identify and combat such associations and activities. Once again however, the persistent (24/7) monitoring of social media feeds would render such an endeavour disproportionate to the ends sought, considering again

---

> The rights to freedom of association and assembly are also threatened by the use of surveillance. These freedoms often require private meetings and communications to allow people to organize in the face of Governments or other powerful actors. Expanded surveillance powers have sometimes led to a "function creep", when police or intelligence agencies have labelled other groups as terrorists in order to allow the use of surveillance powers which were given only for the fight against terrorism.

Similarly, Human Rights Watch (2014), using telecom surveillance in Ethiopia as a case study, found that the Ethiopian State used its surveillance powers to suppress the formation of associations and peaceful assembly.

the implications. Even if the aim was legitimate in itself, it is difficult to say that the impact on privacy would be an acceptable cost, particularly where other alternatives exist (police authorities could for instance conduct existing intelligence collection activities or await reports by the public or media). Situations of emergency, anticipated emergency, or even mere high tension may necessitate such a use of the system for a temporally limited time, particularly with regards to assembly. During an authorised demonstration (particularly where the theme is a contentious one), it may even be justifiable to utilise such a system in order to detect a planned escalation to violence among participants in order to avert emergency. In this scenario, usage would be of a limited temporal scope and not necessarily in order to disband or refuse permission for assembly. It would also empower the appropriate agencies involved to fulfil their positive obligations to protect those practicing their right to freedom of assembly (the system may detect planned violence coming from a counter-demonstration).

7.5 Conclusion

The rationale of this chapter was to define the risks to trust-qualified relations between agents in natural disaster response where a significant variable is the inclusion of systems that collect and process information from social media to assist decisions in natural disaster response, and explore whether there were possibilities for such systems to support trust. This was achieved using IE and trust theory as advanced by Taddeo and Turilli (Taddeo, 2009, 2010a, 2010b, Turilli, Vaccaro and Taddeo, 2010, 2010) which enabled the extension of the concept to include AAs in trust-qualified relations, and utilised Fiduciary Theory to argue that human rights violations undermine trust between the fiduciary state and subject, consequently taking this opportunity to examine the implications of Slándáil-type EMIS for the rights of expression, association, and assembly.

The foregoing analysis demonstrated that trust-qualified relations can be weakened by the presence of inaccurate information, and the extension of use of Slándáil-type systems beyond the domain of natural disaster response, or even increased functionality within the broader domain of natural disaster response. Failures within the Slándáil-type system from all agents involved, and misuse of such systems, can reduce the trustworthiness of any agent implicated within the chain of such failures (ethical and factual failures). This chapter aimed not only to examine and understand the implications posed to trust in this context, but also possible solutions—the one

proposed was the addition of credibility assessment features in Slándáil-type system to reduce the threat of misinformation/disinformation.

Similarly, following from the aim of identifying and analysing adverse (and positive) implications for human rights, the foregoing analysis demonstrated that misuse of such systems can have chilling effects on the rights of expression, association, and assembly, which means that misuse of such systems could effectively deter people from exercising their rights.

The foregoing analysis is of great importance in the overall context of this research, which seeks the positive and practical ends of developing guidelines for the design and deployment of Slándáil-type systems, and it is not without understanding areas of potential harm can useful solutions to such threats be presented, nor can limits to the design and use of such systems be proposed. In identifying with and engaging with issues affecting the ethical and legitimate deployment of such systems here, informed solutions can further be explored in Chapter 9.

8 RESPONSIBILITY AND ACCOUNTABILITY

8.1 Introduction

At an early stage in research, this section was intended to deal exclusively with the value of accountability. As research progressed, it became apparent that no effective analysis of accountability can be complete without the overlapping value of responsibility, as both operate distinctly, yet crucially, towards similar goals and at different levels of analysis, regularly intertwining. While distinct, both values are inseparable in any comprehensive moral analysis, therefore this chapter was adjusted to accommodate analysis in light of both.

Both values serve perhaps the most important role of all those analysed in this research; they are necessary in identifying sources of evil in a moral situation, agents responsible for evil, who should be held to blame, and whether human agents can be blamed at all. Importantly, they concern structures that support the identification and evaluation of moral and morally charged agents. Without responsibility and accountability, it becomes impossible to locate sources of evil and address them with solutions.

Beginning this final analytical chapter will be a brief examination of the challenges to accountability as enumerated in large part by Helen Nissenbaum (1996), imparting to the reader that in the new complex hyperhistorical context, there are factors which contribute to the obscuring of responsibility and accountability.

It will then outline the important distinction to be made between accountability and responsibility within the theoretical framework used, arguing that accountability is a mechanistic construct used to identify causation and subject/object ascription and subsequently address agents that may play causal roles in harm. At a deeper level is responsibility, which is the ascription of object to a morally responsible agent. It will proceed to examine the implications of Slándáil-type systems for the values of accountability and responsibility.

The human rights analysis will begin with an examination of the relevance of accountability and responsibility to Fiduciary Theory, then examining it in practice in the case law of the ECtHR and finally by examining the human rights implications of Slándáil-type systems to the particularly relevant right identified, that is, the right to effective remedy.

8.2 The Problem of Responsibility, Accountability, and Information Systems

A challenge presented by ICTs is the appropriate ascription of responsibility, accountability, and blame in distributed systems where the precise source of an ethical problem (or perhaps the locus of entropy or moral evil) can be difficult to identify, or at least identifying the particular interactions or chain of interactions that result in moral evil. Added to this challenge is the appropriateness (or inappropriateness) of the ascription of responsibility, accountability, or blame to artificial agents or artefacts embedded in such chains of interactions. Superficially, the ascription of socially constructed concepts conceived at least partially to act as deterrents to wrong-doing to artificial agents without intentional states would be problematic (Stahl, 2006a).

The following section will disentangle issues of responsibility, blame, and accountability as they apply (or do not) to AAs with deeper theoretical analysis. Here, it is sufficient to demonstrate the practical problem of identifying the causality of faults in distributed systems (or multi-agent systems). Tracing causality lies at the heart of accountability, which under Bernd Carsten Stahl's (2006a, p.7) account, is concerned with establishing how the relation between subject and object can be verified, "... accountability is the set of mechanisms that allow such tracing of causes, actions, and events."

Tracing causality can be opaque business where networks of agents are involved. Nissenbaum (1996) identifies one challenge to accountability as that of *many hands*. Behind the creation of a computer system are multiple agents with different roles.[172]

Because of the problem of 'many hands', isolating sources of fault can be difficult. An array of persons work on computer systems (or software) and the source of any error or (at worst) moral evil emanating from such computer systems can occur at any point in this network, whether the fault is unintentional or by design.[173]

---

[172] As Nissenbaum (1996, p.28) argues:

> Most computer systems in use today are the products not of single programmers working in isolation but of groups of organizations, typically corporations. These groups, which frequently bring together teams of individuals with a diverse range of skills and varying degrees of expertise, might include designers, engineers, programmers, writers, psychologists, graphic artists, managers, and salespeople.

[173] As argued by Nissenbaum (1996, p. 29):

As Nissenbaum (1996, p. 29) argues, obscured accountability can be either intentional where institutions are arranged to minimise accountability for "negative outcomes", or it can be a by-product of hierarchal organisation where decision makers are "...distantly related to the causal outcome of their decisions."

Just as there are many hands involved in the decision-making behind and design of computer systems, computer and/or software-systems can also be patchworks of different components or modules made by different people, or may incorporate code from earlier versions of systems or different systems, that function collectively towards a given goal (Nissenbaum, 1996). Nissenbaum (1996, p. 30) argues that "...[w]hen systems grow in this way, sometimes reaching huge and complex proportions, there may be no single individual who grasps the whole system or keeps track of all the individuals who have contributed to its various components." It can also be assumed that when computer systems reach such a level of complexity, with various modules/components operating symbiotically and potentially to some degree outside of the knowledge or understanding of the developers involved, tracing the root causes of malfunction or unexpected actions that result in some evil will be difficult. The problem is not just identifying human causality, especially if we assume that all human agents involved did the most they could to produce a functioning and effective system, but also tracing the precise interaction of components or the bad code that causes the undesirable system behaviour. This highlights the next issue presenting an obstacle to accountability in information systems, which is that of *bugs* (Nissenbaum, 1996).

Bugs refer broadly to errors in software, from the modelling to coding of the software (Nissenbaum, 1996). Nissenbaum (1996, p. 32) frames bugs as being an inevitable and endemic aspect of programming, referring to them as "...natural hazards of any substantial system." Faults and errors in software coding can be anticipated yet difficult to find and correct even if they do not result in harmful malfunctions; this is a problem that is exacerbated as a software or computer system is updated or further developed (Birsch, 2004, p. 234). Bugs act as a barrier to accountability as it can be difficult to determine just how inevitable any resultant harms from the bugs could have been

---

> When high level decisions work their way down from boards of directors to managers, from managers to employees, ultimately translating into actions and consequences, the lines that bind a problem to its source may be convoluted and faint. And as a consequence the connection between an outcome and the one who is accountable for it is obscured.

avoided with adequate care and good practice—the question arises, was the software in question exclusively the source of or essential in the perpetuation of a moral evil or was human negligence or mal-intent in the design (or perhaps use) of the system a factor (Nissenbaum, 1996)? This question leads to the next challenge to accountability presented by computer systems, *the computer as scapegoat* (Nissenbaum, 1996).

It can be tempting to blame a computer for any harms emerging from computer systems—they mediate human to human interactions, they perform tasks once performed by humans, and "...human actions are distanced from their causal impacts", with the "...computer's action... a more direct causal antecedent" (Nissenbaum, 1996, p. 34). In view of inevitable bugs, it may not be an unreasonable conclusion that the computer or software is the beginning and end of the source of moral evil, and with qualification, such a conclusion will not be rejected here. However, as the most visible agent in an action that causes harm, the computer or software system presents an obvious target for accountability (to the extent that accountability might apply to AAs) even where human agents may nonetheless be involved and responsible for that harm too (through negligence or mal-intent). Nissenbaum (1996) argues that the computer is cited as the problem (with human agents eluding accountability), either through shirking of responsibility, or through genuine confusion in complex arrangements (the structure and organisation of a multi-agent system) as to where responsibility truly lies. A not unrelated issue is that of *epistemic enslavement*.

Epistemic enslavement occurs where agents occupying an epistemic niche[174] are epistemically dependent on expert information systems (Rooksby, 2009). Van den Hoven (1998, p. 100), as quoted by Rooksby (2009, p.82) defines epistemic enslavement

[174] Rooksby (2009, p. 82 citing Van Den Hoven, 1998) indicates the following criteria as characterising an epistemic niche:

> (i) Inscrutinizability condition: it is impossible to monitor what all the computers in an expert information system are doing (inaccessibility); or to keep track of it all (intractability);
> (ii) Pressure condition: some decisions must be made when there is (a) very little time to make a decision (b) a decision must be made (c) one cannot get expertise from outside the epistemic niche;
> (iii) Error condition: Computers may contain (a) flaws in the specification of the world model of a system (b) brittleness (c) bugs and programming errors (d) limits of testing and proof (e) emergent and unpredictable properties of software, resulting from the interconnecting of systems;
> (iv) Given i, ii, and iii, information systems are inhospitable to the forms of discursive scrutiny by which we traditionally seek to identify experts and to establish reliability on expert opinions [Opacity condition].

as "[i]f a user U is epistemically dependent on expert information system S, and U is narrowly embedded in an epistemic niche of which S is part, then U is epistemically enslaved vis-à-vis S." For the sake of illustration,[175] one can assume that the emergency manager utilising a Slándáil-type system occupies an epistemic niche, and is epistemically dependent[176] on the Slándáil-type EMIS. There is a *prima-facie* case that the emergency manager is epistemically enslaved by the system, particularly where s/he consults the Bonferroni mean aggregation model (or Social Media Index) to consult modelled expert opinion on the risk faced by any given area during natural disaster. The emergency manager is, after all, operating under pressurised conditions where decisions are time sensitive, and the precise modelling of the system cannot quickly be evaluated or assessed. The emergency manager may dispatch assets to a high risk area based upon the modelled expert opinion.[177] Later, upon evaluation it may transpire that the system modelled the data incorrectly through some error, and the area to which assets were deployed did not experience the greatest impacts of the natural disaster. In such a case the temptation may arise to blame the system, without tracing the causality back any further, thus also fulfilling the idea of computer as scapegoat.

A final issue in accountability and information systems (although no claim is made that this section exhaustively documents all challenges to accountability) is that of *ownership without liability* (Nissenbaum, 1996). Nissenbaum (1996, p. 36) argues that "...the trend in the software industry is to demand maximal property protection [of the software product] while denying, to the extent possible, accountability." In many situations, the owner of an object responsible for some harm is typically held accountable for the harm caused by the object (Nissenbaum, 1996). However, in the software industry, through the use of end-user licence agreements (EULAs) for instance, software creators assert ownership of the product whilst pre-emptively denying liability for any harms for which the software is a causal source—the end user is merely a licensee whilst the creator (or whomever holds the intellectual property) maintains ownership, reaping reward while evading risk (Nissenbaum, 1996). Nissenbaum (1996, p. 36) argues that "[t]his trend

[175] And this example admittedly is reductive of the true situation, but a useful abstraction nonetheless—it will be revisited in time.

[176] Rooksby (2009, p.82, citing Hardwig, 1985, p. 338) describes epistemic dependence as a situation "...when one has a good reason to believe true a claim held true by the expert, but cannot assess its truth oneself."

[177] Terms used by interview participant.

creates a vacuum in accountability as compared with other contexts in which a comparable vacuum would be filled by property owners."[178]

These challenges may not emerge independently, they may also converge, thereby making the true locus of blame or accountability yet more obscure.

In what follows, a satisfactory framework for applying accountability and responsibility effectively to computer/information system mediated networks will be parsed out.

8.3 Responsibility, Accountability, and Information Ethics

8.3.1 Accountability and Responsibility

Before proceeding with an examination of accountability, it is instructive to begin with unpacking responsibility in order to come to some distinction between both. Bernd Carsten Stahl (2006a, p. 1) defines and explains responsibility as:

> ...the ascription of an object to a subject. The subject is the entity, usually a person, who is responsible. The object is that which the subject is responsible for. A responsibility ascription thus renders the subject answerable for the object.

Stahl (2006a, 2006b) rightly describes it as a social construct with numerous purposes, both negative and positive, from revenge to retribution, however more generally the aim of responsibility is to improve individuals and social existence, based on the attribution of sanctions depending on whether the outcomes of actions made by responsible persons are undesirable or desirable (or ethical, for that matter). Responsibility "... aims to affect social change for the benefit of those involved in the ascription" (Stahl, 2006a, p. 2). According to Stahl (2006a), responsibility ascription is not complete without an authority or normative basis—or normative rules to which the ascription refers. Stahl (2006a) argues that this is quite simple as regards legal responsibility, but more challenging in reference to moral responsibility. Authority is

[178] Nissebaum (1996, p. 36) goes on to cite (and quote) the example of the Macintosh Reference Manual (1990):

> "Apple makes no warranty or representation, either expressed or implied with respect to software, its quality, performance, merchantability, or fitness for a particular purpose. As a result, this software is sold 'as is,' and you, the purchaser are assuming the entire risk as to its quality and performance." The Apple disclaimer goes on to say, "In no event will Apple be liable for direct, indirect, special, incidental, or consequential damages resulting from any defect in the software documentation, even if advised of the possibility of such damages."

necessary (such as judicial authority for example) in order apply and interpret the normative rules that are the basis of responsibility ascription (Stahl, 2006a). Other dimensions also exist, such as determination of responsibility type (legal or moral), the temporal dimension, and "...the type of ascription (transitive, reflexive, vicarious)... as well as limits and exceptions" (Stahl, 2006a, p. 3).

There are other conditions which must be met for an agent to be considered responsible. According to Stahl (2006b, p. 208),[179] and obviously apparent from the preceding discussion, causality is one of the first conditions, the subject must be linked to the object through a causal chain. Secondly, the responsible agent must have some power of the outcome over the object—if the agent cannot exert any influence over the fate of the object, they cannot be responsible for its fate; full control need not be necessary, however the agent must have at least partial control (2006b, p. 208). Thirdly, the agent must have knowledge, as "[t]he subject must know what is happening in order to influence it" (Stahl, 2006b, p. 208). Fourthly, the agent must be free to act on their knowledge (Stahl, 2006b, p. 208). A final condition, though argued to be controversial by Stahl (2006b, p. 208) is mental states, or intentionality.

What is sketched out here is quite a classical definition of responsibility, is uncontroversial and is broadly reflected in the works of many moral philosophers and ethicists such as Floridi (2013), Birsch (2004), Johnson and Powers (2005), Hellström, (2013), and Noorman (2016).

Problems arise when accountability enters the discussion, as there is a temptation to conflate the two concepts. Indeed, even finding material that examines both concepts separately can be difficult.[180] This is understandable, as both concepts share much in common, however it leads to making the distinction between the two little easier.

[179] The knowledgeable reader might note that the issue of determinism has gone undiscussed here. This is a complex issue that, at worst, strongly undermines the possibility of responsibility, for without autonomy there can be no responsibility. It goes beyond the scope of this research to probe the concept of responsibility any more deeply than what has been done here, and the introduction of discussion of determinism might very well derail the focus of this research. However, it should be uncontroversial for the purposes of this research to assume that all responsible agents, or any agent as defined throughout this thesis, has sufficient options and freedom to decide and act upon those options and, particularly in the case of intentional agents, act upon careful deliberation of the consequences of their actions.

[180] Take a Google search of accountability via Stanford Encyclopaedia of Philosophy for example, where some of the top results retrieved include Collective Responsibility, Blame, and Computing and Moral Responsibility (Noorman, 2016; Tognazzini and Coates, 2016; Smiley, 2017).

Accountability can also be nebulous, sometimes with different researchers attaching different meanings to it, and "...as a result, extant literature treats accountability in a fragmented and inconsistent manner" (Vance, Lowry and Eggett, 2013, p. 10).

This conflation can be somewhat observed in the work of Nissebaum (1996), which while a useful and influential scholarly work, lacks precision in its distinction between accountability and responsibility, at times seeming to collapse the two. The overall substance of the work is unaffected as the issues presented by Nissenbaum are indeed challenges to both the concepts of responsibility and accountability. Here however, recognising that the two differ and can be applied differently and separately, this distinction must clearly be outlined.

Nissenbaum's discussion of the conceptual framework of accountability, blame, and responsibility is undertaken in a single section in her 1996 article *Accountability in a Computerized Society*, and does not so much distinguish between the three concepts so much as blur them together—this is an easy mistake to make, as they will often be used together rather fluidly in identification and evaluation of moral harms and agents, however it is imprecise work. Nissenbaum (1996) refers to accountability as "answerability", without satisfactorily explaining how accountability as "answerability" differs from responsibility as "answerability". Without defining separately responsibility, accountability, and blame, Nissenbaum (1996) proceeds to discuss Joel Feinberg's (1985) framework on moral blame.[181]

Feinberg's framework refers to moral blame, and illustrates where an agent can be deemed morally responsible for a harm that they have caused. As Floridi (2013) argues, blame follows responsibility however not necessarily accountability—this will be

[181] Nissenbaum (1996, p. 28) outlines Feinberg's work as follows:

> Feinberg proposes a set of conditions under which an individual is morally blameworthy for a given harm. Fault and causation are key conditions. Accordingly, a person is morally blameworthy for a harm if: (1) his or her actions caused the harm, or constituted a significant causal factor in bringing about the harm; and (2) his or her actions were "faulty." Feinberg develops the idea of faulty actions to cover actions that are guided by faulty decisions or intentions. This includes actions performed with an intention to hurt someone and actions for which someone fails to reckon adequately with harmful consequences. Included in the second group are reckless and negligent actions. We judge an action reckless if a person engages in it even though he foresees harm as its likely consequence but does nothing to prevent it; we judge it negligent, if he carelessly does not consider probable harmful consequences.

revisited presently. Nissenbaum (1996, p. 28) does admit that the concept of moral blame is separate from accountability, and uses Feinberg's framework to position the analysis of accountability that follows:

> Although moral blame is not identical to accountability, an important correspondence between the two makes the analysis of the former relevant to the study of the latter. An important set of cases in which one may reasonably expect accountability for a harm is that in which an analysis points to an individual (or group of individuals) who are morally blameworthy for it. In these cases at least, moral blameworthiness provides a reasonable standard for answerability and, accordingly, Feinberg's conditions can be used to identify cases in which one would reasonably expect, or judge, that there ought to be accountability.

With such emphasis placed and responsibility and blame and minimal description of accountability, the unique substance of accountability is subsumed. This section will avoid this pitfall and will proceed to outline a distinct definition of accountability.

Already discussed was a partial definition of Stahl's (2006a) definition of accountability, that is, the set of mechanisms that allow a tracing of causality vis-à-vis the connection between subjects and objects, it is concerned with how the relationship "...can be established and verified." Whilst responsibility exists to establish who might be morally praiseworthy or blameworthy in a situation, accountability is the machinery that enable the "...tracing of causes, actions, and events" (Stahl, 2006a, p. 7). Accountability then, is an important part of ascription of subject to object, and a condition of responsibility (Stahl, 2006a). This presents a clear and useful definition of accountability, and one which has been to some degree been replicated across disciplines; for instance, Vance, Lowry and Egget (2013, pp. 10–11) also utilise the definition of accountability as a mechanism that promotes responsibility and blame.[182]

The particular effort here to distinguish accountability from responsibility is made with purpose. They are distinct but interrelated concepts, and in what follows appeals will need to be made to both separately on occasion. Following from the work of Floridi (2013) and as discussed, Stahl (2006a), the goal of accountability is primarily in identification, whilst responsibility is about evaluation. Accountability is a mechanism used to identify those potentially responsible; when examining the actions of the potentially responsible agent we are evaluating them under the criteria of responsibility

[182] Although it might be noted that whilst the researchers recognise the mechanistic aspect, in a reverse problem their conception of accountability does appear to subsume responsibility and blame.

(Were they free to act? Did they have power to act? Did they have the knowledge to act? What were their intentions?). Of course, through accountability processes it is possible to identify sources of evil that cannot be considered responsible moral agents, but may nonetheless be moral agents (artificial agents)—these lack intentionality and cannot be responsible, but nonetheless demand some corrective action.

This problem will be explored in the following subsection, utilising the concepts teased out thus far and exploring others that can help make accountability and responsibility more effective in distributed systems potentially composed of artificial and human agents, where typically they encounter challenges.

8.3.2 Making Accountability and Responsibility Work with Artificial Agents

Artificial agents, as argued, can be potential sources of moral action, that is, they may qualify as moral agents. Certainly even if their actions have no immediate moral impact (fail to pass a moral threshold), their interactions with other agents in a multi-agent system may well yield some impact. Given the problem of inevitable bugs, and the possibility that no human agent can reasonably be found responsible for moral evil arising from an AA, it is important that the AA itself does not elude corrective action following accountability. Without undermining the importance of finding the locus of blame as far as human agents are concerned, it is important to identify all sources or potential sources of entropy in complex networks of agents and to accept that no human agent may very well be to blame.

Whilst AAs cannot be responsible, and cannot be blamed, this is not to say that they do not demand the attention of responsible agents in ensuring, to the best of their ability, that they do not cause harm.[183]

This is a rather foundational aspect of Information Ethics, as discussed in Chapter 2—that non-human agents can be the sources and recipients of moral action, and can thus be identified as moral agents/patients. AAs need to be accounted for, and appropriately responded to, when they have the capacity to contribute to the infosphere in morally

[183] As Floridi (2013, p. 150) argues:

> Since AAs lack a psychological component, we do not blame AAs, for example, but given appropriate circumstances, we can rightly consider them sources of evil, and legitimately re-engineer them to make sure they no longer cause evil. We are not punishing them, anymore than one punishes a river when building higher banks to prevent a flood.

meaningful ways. If the casual contributions to given actions/consequences by AAs can be identified and evaluated, these AAs can be adjusted to function better—either to reduce the entropy that they cause, or improve their output of flourishing or their positive contributions to the infosphere. That non-human agents are subject to accountability should not be considered too controversial or outlandish, so long as one abstains from attributing full-blown responsibility to these agents.[184]

Of course, even though we may identify and acknowledge non-human moral agents that is not to say that there is no bigger picture, that no morally responsible agents are involved, or that responsibility ascription should be shunned. The point is to give due regard to the significant contribution of non-human agents in the infosphere, and not to end analysis of all agents involved prematurely—identifying the computer as a moral agent causing some harm and acknowledging that such a system needs to be re-engineered should not represent the end of the investigation. To expand upon Floridi's (2013, p. 151) example of animals as accountable but not responsible agents, suppose that a rescue dog were to attack a person whom it was supposed to rescue. The dog can be identified as a moral agent, it can be made accountable, but the role of the dog owner in this scenario can also be evaluated. We can ask if the owner knew of its dog's vicious streak, whether its poor treatment of the dog contributed to this streak, and whether the owner used it in search and rescue operations with this knowledge. In this case, there is a morally responsible agent—both the dog and person can be held to account, however only the human is morally responsible, and blameable. Conversely, it may be that the human agent did the best they could in raising the animal, and knew of no vicious streak, and in the dog's long years of service may have had no reason to suspect a vicious streak. In this case, if the dog attacks, it remains an accountable moral

[184] As Floridi (2013, p. 151) argues:

> There is nothing wrong with identifying a dog as the source of a morally good action, hence as an agent playing a crucial role in a moral situation, and therefore as a moral agent. Search-and-rescue dogs are trained to track missing people. They often help save lives, for which they receive much praise and rewards from both their owners and the people they have located, yet this is not the relevant point. Emotionally, people may be very grateful to the animals, but for the dogs it is a game and they cannot be considered morally responsible for their actions. At the same time, the dogs are involved in a moral game as main players, and we rightly identify them as moral agents that may cause good or evil.

agent, though the owner could not necessarily be considered responsible and hence not blameworthy.[185]

Floridi is not alone in his recognition of AAs as sources of moral action worthy of identification and qualified evaluation on their own merits. Stahl (2006b) adopts a similar approach but goes one step farther in arguing for the concept of "quasi-responsibility." Stahl (2006b, p. 211) uses quasi-responsibility to describe the ascription of object to subject where computer systems are concerned, without committing fully to the possibility that computers can be morally responsible.[186]

Stahl (2006b) uses quasi-responsibility in a similar sense as Floridi does accountability; that objects can be ascribed to the AA subject and the AA can be punished where it produces harm to the object, or based on its relation with this object. It is perhaps a red-herring to go this route, to muddle responsibility with accountability (even if acknowledging that it is *not quite* responsibility) when both concepts can satisfactorily serve their role in tracing causation and finding the appropriate loci of blame and the evaluation of agents involved. And blame is important; as Stahl (2006b, p. 211) goes one step farther than Floridi and argues that computers can be blamed within a framework of quasi-responsibility. This is an unnecessary extension of the concept, with hopefully the previous examples having outlined why it is not a reasonable one; one can react to and sanction the vicious and accountable dog, as it were, but to blame it as a responsible agent is ridiculous, and to lay responsibility at its feet means that we risk ending our moral analysis of the situation after identifying the dog as accountable, without reviewing the actions of its human owner.

With the place of AAs in the responsibility, blame, and accountability triad outlined, before concluding this subsection some strategies for better applying responsibility and accountability in complex computer mediated networks will briefly be examined.

---

[185] To quote Floridi (himself quoting Dennet (1996) in the first sentence) for the final time in this research:

...'when HAL kills, who's to blame?'[185] The analysis provided in this chapter enables one to conclude that, since blame follows responsibility, HAL is morally accountable—though not responsible and hence not blameable—if it meets the conditions defining agency... It is responsible and therefore blameable only if, in a science-fiction scenario, it also has a mental and intentional life.

[186] They can merely be in situations that invoke responsibility without meeting all of the criteria such as intentionality, as Stahl (2006, p. 211) argues, "...[i]t is nevertheless useful because it indicates that we are looking at something very similar to responsibility which is nevertheless not quite the same thing as the concept of responsibility we usually encounter".

Based on the work of Stahl (2006a, 2006b) and Nissenbaum (1996), responsibility, that is, ascription of object to subject, should occur as early as possible in a project life-cycle or within an organisational context under the following conditions. Each agent operating within a multi-agent system should be under no mistake about the nature, requirements and limits of the role they have occupied—they should be aware of the goals of their role, the rules governing their behaviour[187] as well as the sanctions they may face for the deviation from their roles or the improper or negligent discharge of their duties. When speaking of responsibility in this way, role responsibility is the explicit model that is referenced (Johnson and Mulvey, 1995; Johnson and Powers, 2005; Stahl, 2006b). Norms surround roles, and such norms must be made explicit to persons occupying those roles. Even beyond the professional context, all agents should be made aware of norms and their ethical responsibilities generally.[188]

Further to this, to ensure effective individual action within professional collective settings, and clear lines of responsibility, following (at least tangentially) the work of Rooksby (2009), herself building on the work of van den Hoven (1998), individuals within groups (and generally as a group) should be assigned meta-task responsibilities. Broadly, meta-task responsibilities refer to an agent ensuring that they are capable of completing a task (internal) and that external conditions are such that they can complete a task (external) (Rooksby, 2009). Without examining the full depths of the meta-task responsibility argument and its applications, or taking any unnecessary detours here, there are important inferences to be made from the concept. For an individual (or group within a complex organisation), to be able to discharge their duties, particularly within sequential work (Bob must complete his task so that Alice can complete hers in its entirety), the individual must be able to ascertain that others have done their tasks, and done so such that they can complete theirs. For this to be so, persons within a complex organisation will logically need to be aware of each-other's responsibilities, or at least appropriate appointed persons will need complete knowledge of who does what along, essentially, a chain of production or interrelated activities. It might be that Bob is assigned some coding on a piece of software that is then passed along to Alice for further and distinct yet dependent coding work. Alice

[187] Therefore clear organisational policy derived from the law and moral knowledge should be promoted.

[188] This is important to note for later, as obviously in moral situations such as those involving Slándáil-type EMIS, agents without professional roles will play morally significant but not professional roles in the situation—that is social media users.

must be able to ensure that Bob completes this task to the extent that hers relies on it (and to the extent that ultimately the entire organisation relies on it) before proceeding.[189] Essentially, an organisation must function such that role responsibilities are known, and different agents are in a position to ensure that the tasks on which theirs depend are being discharged effectively, and that routes are available to ensure that they will be.

This approach should strengthen compliance with rules and best practice, and strengthen accountability (people know their roles, their duties, and the outcomes for which they are responsible, and more importantly others know their peers' responsibilities—*and each person knows that the other knows*). When role responsibilities are clearly defined within an organisation, accountability is made easier, and ultimately moral blame. If a computer system fails, it may be easier to trace the source of the blame. Of course, it may be possible to ascertain that no human agent is to blame if everyone involved discharged their duties to the best of their ability with no knowledge of potential arising harm. In this case, the computer system or AA is accountable for the harm, and this demands response by morally responsible agents. If an AA is the only agent accountable, it remains the responsibility of its creators (or end-users, depending on the exact scenario), to re-engineer it (or destroy it) as appropriate. If corrective measures are not taken, those human agents who are in control of the AA's functionality become morally responsible, and therefore blameable, for the harms it may commit or contribute to.

Knowledge is critical, as *unintentional* ignorance diminishes responsibility, therefore transparency is essential in both responsibility and accountability. Whilst genuine and unwitting ignorance can exonerate, MASes must be designed such that it cannot be blithely used to escape blame.

8.3.3 Slándáil-type EMIS, Responsibility and Accountability

The following will utilise Slándail as a case study in order to understand the implications of Slándáil-type EMIS for the values of accountability and responsibility, structured under the four categories of challenges identified by Nissenbaum (1996). Additionally, a fifth category will be included, that is, whether such technology should be provided

[189] Note that if Alice ignores Bob's work she may be partially responsible for any harm caused, responsibility is scalar (Johnson and Powers, 2005).

open source or closed, a topic which itself has important implications for accountability/responsibility.

8.3.3.1 Many Hands

Many hands were involved in the development of the Slándáil system, and whilst this is apparently natural and intuitive in the development of complex software systems, in the case of Slándáil it may be particularly pronounced. Slándáil was funded under the European Commission's 7th Framework Programme for Research and Technological Development; one of the core purposes of which is to bring together business and research partners from across European borders to work on co-operative research projects in partnership with each other (European Commission, no date). By the nature of its conception and lifespan then it was designed as a collaborative project between many hands across sometimes distant shores, and hence could represent a more extreme example of the issue. However the truth might be that it is quite typical when one considers the nature of modern business; the outsourcing of tasks to overseas vendor companies, which is widely practiced, is arguably analogous in complexity.

The components that form a full configuration of a Slándáil-type system were, as discussed in Chapter 4, developed across numerous states by numerous project partners, and by teams of people within each organisation. Each technological partner worked on their own proprietary software artefacts, sometimes completely external software solutions were incorporated (such as Alchemy API for image analysis), and co-operation and collaboration was necessary between project partners in order to ensure that the components could function symbiotically, and in some cases at a rather deep level of integration.

The particular project, again because of its nature as an EU-FP7 project and the inherent requirement of milestones and deliverables required, benefited from an extensive Description of Work (DoW) where tasks were explicitly assigned to relevant project partners. To this extent, as explicit objects were linked to subjects, there was a measure of accountability. If any individual aspect of a fully operational system, involving components produced by all partners, were to fail then that failure could potentially be isolated to one of a small number of people. If, for example, the proprietary image analysis feature were to fail, the problem could more than likely be traced back to one of few individuals in Ulster University or Trinity College Dublin.

Matters threaten to become more opaque as deeper integration between separately developed, standalone, components occurs, such as a deeper integration between the Social Media Monitor component, for instance, and SIGE. At this stage however, individuals will still have been formally delegated the task of integration, and if any errors arise as a consequence of integration, or are suspected of having arisen as a result of integration, the individuals responsible for this task can be identified and evaluated in their performance.

Even with subjects and objects clearly ascribed within an organisational structure, that is not to say that there will be no challenges. In the case of Slándáil, partners remain distant in geographical terms. This intuitively stands as an impediment to clear knowledge and understanding of the work of peers, exacerbated by the differing specialities of the individuals represented by different partner organisations (linguistics, emergency management software development, and text and image analytics). There can be no single individual with a comprehensive knowledge of the inputs and processes involved, only individuals with overlapping knowledge. There are many hands, and in many different places, and any lack of communication on work that has been done or that needs to be complete may yet lead to oversights in completion of necessary tasks where certain items of work may require closer collaboration between individuals in different partner organisations. The clear and precise ascription of subject to object may not always be clear in fluid and dynamic environments, task responsibilities may not always be precisely formalised or be outlined in a single document (such as a DoW) and tasks may arise in an *ad hoc* or improvised manner.

Locating the causation of any harms arising from these oversights—in areas of converging interests and expertise, for example—may provide a challenge to accountability insofar as causality may be obscured.

Reliance on third party software, such as Alchemy API in the Slándáil project, also presents problems. Whilst difficult to pinpoint moral harm which might arise from Alchemy in particular, if it plays a causal role in any significant failure, the product was not developed within the framework of the specific project and was licensed from a third party (the parent company is IBM). In this case, whilst it might be possible to trace causation to Alchemy, and perhaps even establish responsibility within the framework of the project (by identifying individuals who worked directly with or authorised its inclusion), the organisational structure of Alchemy's development team and

management is external to the project and there may not be defined lines of ascriptions of subject and object—accountability is not immediately available, and precise causal links are obscure.

Of a lesser order challenge than this is the further development of pre-existing software artefacts that existed before project conception. Take CiCui for example, in this case. CiCui pre-exists the start of the Slándáil project, and has been built on and otherwise adapted since. The most recent iteration of the SMM in particular uses code from CiCui, but is a separate, streamlined entity built with additional code. In such cases (though not necessarily here), there may be a limited understanding of the technology's pre-existing faults and limitations, and lines of causation and responsibility may be obscured by time and changes in staff.

8.3.3.2 Bugs

Not entirely dissimilar from the previous challenge is the issue of bugs. The issue of many hands may exacerbate the likelihood of bug occurrence, tracing, and correction, as many different software artefacts may operate in tandem or with some degree of symbiosis depending on their level of integration. The issues of bugs and many hands are evidently overlapping.

During the interview stage, participants were not asked about bugs or the bug-testing process. Nonetheless, logical inferences can be made from available information. In the case of the Slándáil project, there are many different software artefacts, each one capable of suffering from a bug, and the individual artefact or system component precisely at error may be difficult to trace without adequate error logs, particularly where there is a larger degree of integration. Bugs can have unpredictable consequences, and in the sensitive moral situations to which Slándáil-type EMIS apply, they could be anywhere from negligible to even quite severe. It can be presumed that consequences might be quite light,[190] to moderate,[191] or even severe.[192]

[190] A malfunction results in showing emergency managers noise in addition to signals, which would reduce the efficiency of acquiring situational awareness from social media feeds but would not necessarily critically jeopardise emergency management and response.

[191] A critical failure that perhaps causes the Social Media Monitor to stop functioning—valuable information might be missed but responsible emergency managers would nonetheless have access to other information sources.

[192] Suppose that a bug in any system component providing an expert opinion, the Bonferroni mean aggregation model for instance, caused it to provide inaccurate risk-related information that may not be immediately verifiable and prompts misallocation of resources.

As argued, with task or general role responsibilities outlined, it may be possible to trace causality back to individuals with the responsibility for the component or artefact that failed or caused harm, though it may not always be easy, and the individuals whose contribution should be evaluated may not ultimately be responsible for any harm caused.

There is obviously need for robust bug testing before any software is deployed in an emergency setting. Failure for appropriate precautions being taken in testing and bug identification could at worst place lives in danger in such sensitive conditions under which a Slándáil-type EMIS would operate. Without such processes and procedures in place, the individual organisations tasked with the development of each software component or artefact would be accountable for any harm caused through negligence. Collectively, arguably all organisations tasked with the design of system components that must function together effectively would be accountable for not ensuring, to the best of their ability, a system that functioned in its totality bug free. In a collaborative project where systemic and structural failures are the cause of harm, the collective entity, or multi-agent system, must be held to account, as well as its individual parts that are in leadership roles, as well as held responsible and therefore blameable as appropriate (principal investigators and persons tasked with quality assurance in the case of Slándáil would be of particular interest).

Of course, where satisfactory rules and procedures are in place for identifying bugs and all individuals have done the best they can to identify them, and in spite of an excellent quality assurance framework, critical bugs still manifest upon deployment (which may result in harm), these human agents cannot be said to be responsible.[193] Here, the Slándáil-type EMIS (or a component thereof) is accountable exclusively, however both its creators and its end-users remain responsible for the system and its uses, and the fault must be corrected by the appropriate personnel, and if not, the system or individual component causing the fault decommissioned. While AAs can be accountable, their creators and users depending on the circumstances remain responsible for their use, and knowledge of their flaws compels them to either put such AAs into disuse or correct/re-engineer them before using them again.

[193] Again, presuming every reasonable effort was made to mitigate bugs to the hypothetical bug had not manifested before, or under the same conditions.

8.3.3.3 The Computer as Scapegoat

When AAs play such a pivotal role in information collection, analysis and output in highly pressurised, time sensitive situations with a need for highly accurate information which will inform decisions in disaster response, it can of course be a great temptation to scapegoat the AA decision support system for any erroneous information or system failure that results in lost time and resources.

At the deployment phase during a live event, there are a great number of agents involved in the moral situation, composing the multi-agent system (of both agents and patients) along within the network and causal chain; there are the private or scientific institutions that developed the system, the emergency management and response agencies, the technological artefacts, social media users and disaster affected (these latter categories likely overlapping substantially). The system, Slándáil, mediates between emergency managers and disaster affected/social media users, and will play a role in determining the actions taken by emergency managers towards the relief of disaster survivors. The relations between all of these agents, and the potential outcomes, can be complex, as described in the preceding chapters (particularly in Chapter 7).

Causation of harm can be difficult to identify in such complexly interwoven relationships between agents (or agents and patients). The system is tasked with presenting accurate and timely information to emergency managers, but the system's output is only as reliable as the information it collects from social media (or in combination with other administrative or academic sources as the case may be with the Bonferroni model). Where the system presents inaccurate information, does the fault lie with the system, its creators, social media users, or farther down the line where misallocation of resources has occurred, the emergency manager?

In any eventuality, the responsibility is to some degree distributed throughout all of these agents (with the exception of AAs, bereft of responsibility but nonetheless expected to function ethically and optimally) in order to ensure the most positive outcomes. There is a limit to which the AA can be scapegoated, and the previous chapter did much of the work in demonstrating this. The example of misinformation/disinformation during disaster situations is a useful one to turn to in order to demonstrate how responsibility is distributed. Misinformation and disinformation arising from human agents (social media users or disaster affected) can

compromise an emergency manager's capacity to make effective decisions, and therefore where-ever a human agent knowingly posts untrue information that has the potential to mislead the response agencies, this human agent is partially responsible for any harms arising from the potential misallocation of resources that occurs, and potentially at worst any deaths that might occur as a result of this misallocation of resources. And yet in this situation the emergency manager may be, and probably is, morally blameworthy too, should they have made insufficient efforts towards confirming the veracity of reports which they took to be true at face value.

The social media user as a morally blameworthy agent itself exposes yet further challenges to accountability in the moral situation. The social media user may be operating anonymously, with a fake name, pseudonym and potentially a spoofed geo-location. For the social media user to truly be accountable or held responsible and sanctioned appropriately, they need to be identifiable. Law enforcement organisations may request personal data from service providers such as Twitter, but even this may not yield results if the user registered to the service with fake details and used methods such as a VPN to obscure their true IP address.[194]

To what degree does the AA play a causal role in any resulting harms? In this particular scenario,[195] as it fails to correctly classify true and relevant information, it is shown up as ineffective at the task with which it was delegated—it is implicated as accountable and should be patched or otherwise updated until it can be trusted to perform to within an acceptable margin of error.

If the creators and designers license the use of a system knowing that it would be ineffective, without declaring its limitations so that it was well known that emergency managers should exercise caution in acting upon the information that it presented, they are partially responsible and blameable for any resulting harm caused. The previous chapter highlighted that there are potential solutions for automatically assessing the credibility of information on social media. This, paired with the demonstrable

[194] This is therefore a structural issue within social media and not one that will be argued to be bad in itself. It is beyond the scope of the present research to probe the problem further, though while it may be noted that this anonymity challenges accountability (the subject to which an object has been ascribed is unidentifiable beyond a superficial level), there are benefits, for instance the promotion of freedom of expression or speech in dangerous regimes that punish subversive opinions.

[195] Assuming that it presents a considerable number of inaccurate messages, given that such systems cannot reasonably be expected to perform with 100 percent accuracy.

prevalence of misinformation/disinformation present on social media, makes inclusion of such solutions (to the extent that they are compatible with the system) a duty incumbent upon the creators/designers. Where there is a known problem, and a practical (if partial assuming that the solution is not 100 percent effective) solution available, and the creators/designers have the power (knowledge of the problem and solution should be assumed to be known by professionals) to implement the solution, there is an arguable duty to do so, or at the very least to try. Licensing the product for use without having attempted to mitigate the problem of mis/disinformation with an available technical solution makes the creators/designers partially morally responsible for any arising harm, though less (yet not still completely without blame) responsible if the limitations of the system have been appropriately outlined to the end-users. If the creators/designers can demonstrate that no solution was compatible with the system, or reliable, then this would cease to be an issue. The burden of proof would be on the creators/designers. Again, generally, if the creators/designers knowingly release a product with known and potentially dangerous flaws, they are morally responsible and blameworthy for any harms that arise as a result of use of the system.

Assuming that designers/creators release a product that was thought to run effectively and without known fault, observed the highest professional standards in its creation and tested it rigorously for bugs, then the system itself may be the largest causal factor in any harms arising (in a situation, at least, not involving misinformation/disinformation—assume full system failure and inoperability). The system is accountable, however this accountability then demands responsibility, as argued earlier. It must be updated by the creators/designers or decommissioned by the emergency managers. Failure to act on this will make both the creators/designers (assuming that failure event has been reported to them) and emergency managers responsible for harms arising from future failure events.

In many cases to hold the AA accountable will not be appropriate, and an investigation into causality involving all active agents in the moral situation will be required. Many potentially harmful uses of systems such as Slándáil cannot plausibly rest exclusively on the system itself, as preceding chapters have shown.[196]

[196] Misuse of personal data and privacy intrusion will often be caused by human agents, by and large, by persons acting outside of ethical and legal boundaries, and the same can be said for use of the system that unfairly privileges persons in society who are already privileged, and marginalizes further the already marginalized. Human agents, that is emergency responders, will

This last point again raises the issue of epistemic enslavement. The computer may be the scapegoat where the emergency manager simply states, "I took this course of action based on the information available to me as presented by the Slándáil Social Media Monitor," or in the case of an operational Bonferronni model, "I took this action based on the expert opinion provided to me by the system." Due to the plurality of resources with which emergency managers work, they cannot reasonably cite epistemic enslavement to indemnify them in *most* cases.[197] They will have numerous means to verify information that is received, they (as noted by Rooksby, 2009) are autonomous, they have choices, and their decisions are not ultimately bound by the information displayed by the system. In the case where they act upon information reported by social media users, again the system is not to blame, but both the emergency manager (who may have had alternative information sources or the ability to verify the reported information) and social media users. Where the system models expert opinion using risk analysis, and can thereby essentially form a hierarchy of areas requiring resources based on risk, the waters become a bit more muddied, as no alternative source of information may be immediately available and as the information produced is numerical, verifying risk on an area by area basis may not be easy or possible using real world assets or alternative data. In this case the problem is more structural, as the emergency management agency as a whole failed to conduct thorough risk assessments ahead of the occurrence of a disaster. The individual or small group of emergency managers are not necessarily responsible for moral harm acting on the basis of the only information available to them, though the agency as a whole is ultimately accountable for inadequate preparation. The role of a Slándáil-type EMIS should be to provide an additional stream of information for consideration prior to action, not to replace pre-existing methods of information collection, or substitute for methods which should be in place but are not. Nevertheless, should a situation arise where alternative sources of information are absent and reasonable means of confirming information presented by a Slándáil-type EMIS are not available, the emergency manager is not responsible for acting on the information provided by the system if in their estimation it is plausible, although this does not necessarily indemnify the agency as a whole if the lack of available alternatives and capacity for verification was a result of negligent

---

also be likely negligent where their decisions, informed by Slándáil-type EMIS, are not sufficiently corroborated by additional sources of information or verified by some form of additional investigation.

[197] Though it cannot be rejected that this is never a valid excuse to some degree.

organisational failure. If automated expert opinion, which gives unexpected inaccurate results, leads the epistemically enslaved emergency manager to action, then a significant portion of the fault lies at the system and leaves at least the individual innocent.[198]

Technical solutions can to some degree improve accountability, both of human and artificial agents. Previous Slándáil ethical research noted some such solutions as restricted user access and journaling systems and will not be contradicted here (Jackson, Aldrovandi and Hayes, 2015). As they operate remotely at least, Topic Analyst, SIGE, and the Social Media Monitor are password protected and therefore improve the transparency of individuals accessing the programmes and who are therefore potentially responsible for actions taken with the systems or based on system output. Today's technology is powerful, and comprehensive digital records can isolate user actions on the system. One interview participant from the technology stream suggested that monitoring of end-user behaviour could be as granular as key-stroke logging, although it is unclear if any artefacts studied utilise this method. It is also incumbent that such systems have the capacity to record the dynamics of internal code when operational, in order to enable the identification of bugs, and offer bug reporting mechanisms for end-users so that they may be corrected and in order to potentially exonerate human agents of responsibility for harm where the system was at fault. Finally, such systems will need to archive social media messages for a limited period of time in order to provide documentary evidence of information that led to decisions made by emergency managers, with a view to the destructive capacity of mis/disinformation by social media users and to hold them responsible insofar as this is possible or appropriate.

8.3.3.4 Ownership without Liability

It can be presumed that software creators/designers seeking to develop Slándáil-type EMIS now and in the future will often wish to retain some degree of control over their products[199] and also seek to mitigate against potential claims based on the failure of their products through end user license agreements (EULAs). The Slándáil project has been no exception (with some qualifications), and one legal deliverable in particular

[198] Although the calculations of the Bonferroni aggregate model, to the extent it weights false social media reports in its analysis and suffers from no bugs, still means that human agents—social media users—remain responsible.

[199] With the possible exception of institutions or individuals who adhere to open source philosophies.

(*D2.6 Licence for the Use of a Disaster Management System*) provides a useful case study of the challenge of ownership without liability.

Attached in the referenced deliverable is a model (therefore not necessarily final) licence agreement for the Slándáil system which outlines the responsibilities of end-users (licensees) and the creators/developers/intellectual property holder (licensors) as well as the limitations of responsibilities, particularly of the licensor (Corbet *et al.*, 2017, pp. 6–26).

A cursory inspection of the clauses of the model licence agreement reveals in C*lause 10 Limitations of Liability* a reasonable position to minimise liability; there is an effort to mitigate against potential claims, or at least moderate the scale of claims between licensee and licenser—it is reasonable as the release from liability is within the boundaries of applicable law (Corbet *et al.*, 2017, pp. 10–11):

> Except as provided for under clause 3.2 clause 8, neither Party shall be liable for any loss, damage, costs or expenses of any nature whatsoever incurred or suffered by the other Party that is (a) of an indirect, special or consequential nature or (b) any loss of profits, revenue, data, business opportunity or goodwill.
>
> To the extent that either of the Parties has any liability in contract, tort (including negligence), or otherwise under or in connection with this Agreement, including any liability for breach of warranty, their liability shall be limited to [the Licence Fee].
>
> Nothing in this Agreement excludes or limits either Party's liability for:
>
> i. death or personal injury resulting from its negligence or the negligence of its employees or agents; or
> ii. fraud or fraudulent misrepresentation; or
> iii. its obligations under clause 8 (Indemnity); or
> iv. matters for which liability cannot be excluded or limited under applicable law.

This effort to avoid liability is couched in broad terms, and limits the liability of licensor to licensee to the fee for using the system. It is not however an absolute effort of shirking liability, as can be read in provisions i-iv. A footnote also states that "Note: limitation of liability (if any) to be negotiated between the parties", which does provide for the possibility that the clause is tentative (Corbet *et al.*, 2017, p. 10).

On further inspection, extending past the concept of liability (to the extent that it applies mostly to duty to compensate financially for damages caused), the model licence agreement does attempt to indemnify the licensor, and transfer responsibility to the licensee, as can be noted from *Clause 8 Indemnity* (Corbet *et al.*, 2017, pp. 9–10):

The Licensee shall indemnify the Licensor against all Claims brought against the Licensor which relates to or is caused by:

a) any decision or action taken by the Licensee (and any consequences that flow directly or indirectly from any such decision or action) based wholly or partly on the Software; or
b) any damage to, or loss of, life or property caused from the Licensee ordering an evacuation (or not ordering an evacuation) based wholly or partly on the Software; or
c) any breach by the Licensee of any laws or regulations applicable in the Territory, including, but not limited to, the laws of contract, data protection, privacy, copyright and human rights laws; or
d) the failure by the Licensee to secure all necessary consents and permissions required in order to lawfully use the Social Media Data in the Territory for the Purpose.

Here, the licensor seeks unequivocally to assert indemnity against any actions taken by the licensee that may cause harm and which may have been based on information presented by the system. This is certainly a challenge to accountability if the licensor is evading the possibility that its system can play a causal role in any harm caused by an emergency manager's decision based on information received, given that the licensor has a duty to ensure that, to the best of their knowledge and ability, the system produces reliable information and any dereliction of duty to ensure this should render them accountable and appropriately sanctioned commensurate with their degree of responsibility. That is to say, if negligence on the part of the licensor causes a fault in the system's effective operation that jeopardises emergency response, the licensor must accept and not attempt to evade responsibility.

Additionally, the preceding terms and conditions are reflected in clause *3.2 Conditions of Licence* (Corbet *et al.*, 2017, p. 8):

The Licensee acknowledges and accepts the following:

a) That the Licensee uses the Software entirely at their own risk;
b) That the Licensor does not create, edit, own, moderate or otherwise control the Social Media Data that is harvested via the Software;
c) That the Licensor does not guarantee or warrant that the Social Media Data is accurate or complete and expressly disclaims all liability for any loss or damage resulting from your reliance on the Software;
d) That the Licensee is entirely liable for any liabilities, damages, losses, costs, fees and any other expenses incurred as a result of using the Software and, in accordance with clause 8, shall indemnify the Licensor against any such losses incurred by it as a result of your use of the Software.

Here it is stipulated that the system is used at the risk of the licensor, and the point is reasonably made that social media data can neither be guaranteed to be complete nor

accurate, yet mis/disinformation is not the only risk, but system malfunction arising from negligence.

Exacerbating these challenges to accountability is *Clause 13.11 Announcements* (Corbet *et al.*, 2017, p. 14):

> *Announcements.* Neither Party shall make any press or other public announcement concerning any aspect of this Agreement, or make any use of the name of the other Party in connection with or in consequence of this Agreement, without the prior written consent of the other Party.

This clause in particular is concerning, especially with regard to a system utilised in the context of public service and in view of it being contrary to the value of transparency, and in so doing obscuring accountability. The clause does allow for publicity of the licence agreement where consent is obtained, though as the system operates in and for the public service/interest, in a democratic society, and in the interest of accountability, details of such transactions and the division of responsibilities between parties should be known to the public regardless of the consent (or lack thereof) of either party—particularly as the public should be entitled to challenge the terms and conditions of such agreements; they involve public, democratic institutions and public expenditure (licence fees). As the licence agreement asserts licensor ownership and has important indemnity and limited liability clauses, it is especially important that members of the public should be made aware so that these terms and conditions may be challenged. Accountability cannot be facilitated easily where agreements involving public institutions are covert, where terms and conditions would attempt to deflect responsibility disproportionately to one party (the licensee), where the public (potentially social media users and/or disaster affected) is unaware of these conditions and unaware that their ability to individually hold the licensor accountable might be obstructed by unethical and responsibility evasive terms and conditions.[200]

As a caveat to this critique, it must be noted that the re-quoted clause that follows must be weighed against all clauses that assert indemnity (Corbet *et al.*, 2017, p. 11):

> Nothing in this Agreement excludes or limits either Party's liability for:
>
> i. death or personal injury resulting from its negligence or the negligence of its employees or agents; or

[200] Though it should be noted that no written agreement supersedes the law itself, and it can probably be assumed that justice would prevail within the courts irrespective of the wording of a written contract.

ii. fraud or fraudulent misrepresentation; or
iii. its obligations under clause 8 (Indemnity); or
iv. matters for which liability cannot be excluded or limited under applicable law.

The presence of this clause fundamentally accepts that the licensor can under certain conditions be held accountable, yet on balance the agreement takes any opportunity to minimise the situations where the licensor is accountable, a fact which is unfair to the licensee and the public, potential moral patients, who may seek redress from and/or sanctions against the licensee where negligence can be identified.

Another caveat is that (and as is acknowledged in the introductory section of the document), there is a clause that allows for amendment of the agreement (Corbet *et al.*, 2017, pp. 3, 12). This means that the agreement would not be monolithic, and could be subject to revision or negotiation.

As to the particularities of ownership, the agreement asserts this robustly and prohibits altering of the system or reverse engineering of it by licensee except where this is permissible by the law (for error fixing or issues of interoperability) (Corbet *et al.*, 2017, p. 4). This is provided for under C*lause 4 Supply of Software* (Corbet *et al.*, 2017, p. 8):

> a) Except as expressly permitted by this Agreement, the Licensee shall not modify, adapt, disassemble, reverse engineer, decompile, translate, or otherwise attempt to discover the source code of the Software or permit any of these things to happen, except as expressly authorised by applicable, mandatory law governing the rights of software licensees.

Part b) of the same clause goes on to say (Corbet *et al.*, 2017, p. 9):

> The Software is provided “as is” and the Licensor shall have no obligation to upgrade, bug-fix, provide support or maintenance services, or provide any information, assistance or consultancy in relation to the Software, unless agreed between the Parties.

Here the licensor asserts no obligation to provide ongoing support or maintenance of the system, which is yet another flaw in the agreement. The licensor is responsible for providing a working and effective product, and is responsible for fixing any errors of their own making to the extent that they can be fixed (if not, if sufficiently serious, the licensee should decommission the system), and must provide any information as required that facilitates the effective operation of a system that is used under such sensitive circumstances. The introductory section of the deliverable does acknowledge

that law requires the licensor to provide these services, therefore it seems to be redundant and unenforceable (Corbet *et al.*, 2017, p. 4)

The assertion of ownership and prohibition of altering the system by the licensee by the licensor (insofar as the law can prevent this) is not inherently bad. This clause can theoretically prevent the licensee from modifying the system such that it can be applied to a wider range of events or non-emergencies (though note that as the system is highly customisable, no in-depth re-engineering should be necessary to achieve this goal). This, taken in conjunction with the very specific set of conditions under which the licensee is licensed to use the system militates against the possibility of undesirable function creep (or conversely, admittedly, any additional desirable functionality for a wider scope of events). The agreement states (Corbet *et al.*, 2017, p. 7):

> The Software can be lawfully and usefully deployed to assist civil protection and emergency response agencies in the EU to prevent, manage and respond to natural and man-made disasters.

It additionally provides for termination of licence where the agreement is breached by the licensee and the possibility of sanction being initiated by the licensor to the licensee, for example, under *Clause 7.1* of *Clause 7 Infringement of Intellectual Property Rights* (Corbet *et al.*, 2017, p. 7):

> *Infringement of the Software.* The Licensee shall inform the Licensor promptly if it becomes aware of any third party infringement, or potential infringement, of the Software. The Licensor shall have the exclusive right to determine whether or not any litigation shall be instituted or other action taken in connection with any infringement, or potential infringement, of the Software.

And *Clause 11.3 (i)* of *Clause 11 Duration and Termination* (Corbet *et al.*, 2017, p. 11):

> Either Party may terminate this Agreement at any time by notice in writing to the other Party (the "**Other Party**"), such termination to take effect as specified in the notice:
>
> if the Other Party is in material breach of this Agreement and, in the case of a breach capable of remedy within ninety (90) days, the breach is not remedied within ninety (90) days of the Other Party receiving notice specifying the breach and requiring its remedy...

To this extent, the agreement functions as a mechanism of accountability that, in theory, limits the use of the system by the licensee and establishes the right of the licensor to withdraw service in the event of misuse of the system. Whilst the licence agreement largely disproportionately places responsibility on the licensee as the licensor evades it, it is positive and desirable that the licensor has the power to police use of the

system and terminate the service in the event of system misuse. The licensor is responsible for its creation, and is obliged to the extent of the power it has to do so, to ensure its ethical use.

The agreement also requires that the licensee comply with the laws of their jurisdiction in the use of the system including data protection laws (Corbet *et al.*, 2017). The deliverable also features a legal check list intended to be read before the licensee signs the agreement. This checklist was the culmination of legal and ethical research undertaken during the project, and was designed in order to identify (though not necessarily exhaustively) a range of ethical and legal issues with which the licensee should be familiar, including under the categories of data protection, privacy, copyright, and human rights (Corbet *et al.*, 2017, pp. 17–26). This further establishes the responsibility of the licensee, and provides some normative rules against which the licensee can be held—it aids the licensee in being aware of their legal obligations, and importantly it signals that the *licensor knows what the licensee's legal obligations are*. The breach of any element of the checklist could count as grounds under which to terminate the licence, and contributes towards the agreement being an mechanism of accountability to an extent, even if the licensor holds near unilateral power over the licensee in the current draft of the model licence.

The Model License Agreement is a cautionary tale to add to Nissenbaum's (1996) anecdote about Apple's EULA. Whilst perhaps not as extreme, it does have the effect of attempting to deflect accountability from the licensor to the greatest extent possible. In this regard, it is a warning about challenges to emerge as Slándáil-type EMIS become more commonplace as other creators/licensors may well attempt to craft similar agreements. This would be the wrong approach to take. Morally responsible agents such as creators/licensors must accept their share of the blame where it is due and face any sanctions applicable for their actions (or omissions) that can be deemed harmful. Deflection undermines accountability, and accountability plays an important role in the correction of error (or evil).

A positive lesson to take from this licence agreement however is the leveraging of ownership for control towards positive ends. The agreement provides legal and ethical information to assist licensees in responsible system use, and also requires compliance with national and international law by the licensee. The agreement also explicitly states that the system may only be used for natural and man-made disaster, thereby limiting

the possibility of function creep. The licence can be terminated where licensee use is illegal, therefore it does function as a limited mechanism of accountability in itself and shows that ownership (by a responsible entity) can serve a valuable purpose in terms of acting as one check and balance.

8.3.3.5 Open or Closed Source?

The previous subsection raises an important issue, which is whether it is ethically better for Slándáil-type systems to be provided under open-source (that is, with limited restrictions to modification, usage, and distribution) or closed-source (that is, with firm restrictions upon usage, modification and distribution) licences. In the preceding subsection, the closed-source model was argued as being ethically favourable in the particular case. This subsection will explore the matter in more depth, with a particular concern for the values of responsibility and accountability.

Describing, essentially, the difference between open- and closed-source (instead using the terms "free" and "proprietary"), Chopra and Dexter (Chopra and Dexter, 2009, p. 287) state that:

> The fundamental difference between free and proprietary software lies in the nature of the actions that users of the software are permitted to take. Proprietary software, relying on trade secret, licensing, and copyright law, restricts user actions via end user license agreements (EULAs); free software licenses eliminate, to varying degrees, restrictions on user actions. The difference between proprietary and free software, as established by software licenses, is not a question of price. A free software package may cost as much as a proprietary package; that is, "free" only affects what the user may do with it once she has procured it.

The software freedoms typifying this free or open source philosophy are (Chopra and Dexter, 2009, p. 288):

- The freedom to run the program, for any purpose (freedom 0)
- The freedom to study how the program works, and adapt it to your needs (freedom 1). Access to source code is a precondition for this.
- The freedom to redistribute copies so you can help your neighbour (freedom 2)
- The freedom to improve the program, and release your improvements to the public, so that the whole community benefits (freedom 3). Access to the source code is a precondition for this.

The free philosophy is clearly very attractive, it democratises (or socialises) ownership and innovation, and it challenges monopolies (especially where "free" means "at no

cost") that can harm consumers. Consider LibreOffice,[201] a free office software suite with much the same functionality as Microsoft Office, or Ubuntu,[202] a free alternative operating system to Microsoft Windows. They give consumers choice, and low resource entrepreneurs and developers the option to participate in the market.

Aware of the potential for harm of the Freedom Zero philosophy, Chopra and Dexter (2009, p. 290) bring to attention the scientist's dilemma; "...should I allow others to use the knowledge I have produced, knowing as I do that it may be used for morally questionable ends?"

Broadly, Chopra and Dexter's (2009) ultimate answer is a hesitant no, acknowledging the dangers of restricting the dissemination of knowledge and the role that scientists typically already do play in this (scientists have the responsibility to, and do, publicise and bring about debate on the issues involved). The thrust of the argument in favour of Freedom Zero is that restrictive licences can stymie scientific progress (moral values clash, and a creator may prohibit licence to use their creation in what they may deem an immoral context, yet the prospective licensee may consider their work ethically justifiable) and that Freedom Zero "...supports deliberative discourse within the development and user communities" (Chopra and Dexter, 2009, p. 294).[203]

Chopra and Dexter (2009, p. 295) do however retreat to the position that:

> A scientist/programmer is justified in placing substantive legal restrictions on the use of knowledge/programs created by him when a morally objectionable use of the work in question can be anticipated. If no such use can be anticipated then the scientist is justified in releasing this knowledge for the untrammelled use by everyone. No moral approbation should be attached to the release.

[201] See https://www.libreoffice.org/
[202] See https://www.ubuntu.com/
[203] On the subject of discourse, Chopra and Dexter (2009, p. 294) argue:

> If an owner or creator were able unilaterally to forbid a particular use of some licensed software it would limit opportunity for a rich discussion and concomitant education about the moral dimensions of technology... It is not only the inventors or discoverers of an ethically charged idea or object that are invested in its fate: individuals, who may be benefited or harmed by it, as well as society, may legitimately stake a claim in the discussion about its uses. If so, discussions about possible uses and bans on them involve the entire community and invite the broadest deliberation and discussion. As the contentious cloning and stem-cell debates demonstrate, all the stakeholders in a discussion may need to be identified and engaged before any decisions can be made about research agendas and policies.

This latter position is the one that will be taken here. Maximally open-source or free software licence agreements serve an important role in society, in scientific discovery and innovation, and the democratisation of knowledge and its use for the masses—the concept as a whole will not be rejected. Where the possible negative applications of a technology are ambiguous or uncertain, open knowledge sharing and its facilitated public discourse are desirable—here, not only scientific progress can take place but also development of our shared understanding of values and ethics, particularly as they relate to technology. Other times, however, morally objectionable use of knowledge or technology might be more readily apparent. To bring the discussion more explicitly to responsibility, the scientist/developer who created the technology, with knowledge of its potential, if not likely, evil use is obliged to try to—within the best of their ability—prevent this evil use. Assertion of control over their creation, and its knowledge, through closed-source type licences enables them to prevent their knowledge or technology from falling into the wrong hands.

For one agent to pass along to another some object that can do harm, with knowledge that the receiving agent would do harm with it renders the gifter almost as responsible for the resulting harm—they had the knowledge and power to stop it.[204]

In the case of Slándáil and Slándáil-type systems, hopefully the preceding chapter in particular did much to outline the potential evil applications of the system by malicious agents. In light of this, for a creator to widely disseminate the source-code so that it could be engineered and used indiscriminately would be morally wrong, and the creator would be morally responsible for resulting harms.[205]

The closed-source type licence then is preferable in the case of a Slándáil-type system. It allows the creators to exert some control over its use, to be selective as regards to whom it is licensed, and it allows them to terminate the agreement where misuse is detected. It enables them simply to be responsible for their creation. In the case of

[204] Consider the following, If Bob were to lend a handgun to Alice, knowing that she had a history of violence and was banned from purchasing arms, we would hold Bob partially responsible and morally blameworthy were Alice to subsequently shoot her neighbour.

[205] Just as we would blame Bob for giving dangerous Alice her murder weapon, we too would blame the creators or IP holders of a Slándáil-type system if they were to make it available to the authorities of rogue regimes that we might expect, with some degree of certainty, would attempt to use it to monitor and suppress legitimate protests—and we might blame them even moreso should they explicitly license the system on a closed source based license to such a regime, knowing the likely evils for which it would be used.

Slándáil's Model Licence Agreement, there are sufficient clauses to indicate intended responsible use, chiefly being that it is licensed for use in the EU (Corbet *et al.*, 2017), a region that benefits in particular from strong data-protection regulations, a strong human rights regime, and relatively good respect for the rule of law. The Model Licence Agreement also enforces accountability, by suggesting familiarity with applicable law and requiring compliance—this provides normative rules allowing the licensor to identify the licensee as a cause of harm, a contractual deviant, and to act upon this through agreement termination. It improves accountability of the licensor in scenarios where they provide known malicious agents with a licence, it provides a paper trail of evidence that the licensor provided the system to a known malicious agent contrary to its own stipulated terms and conditions, or allowed this agent to continue use despite misuse of the system.

It has been shown that the Model License in question attempts to evade licensor responsibility and accountability. This does not mean that the underlying concept of a closed-source licence is bad, merely that it needs to be revised such that accountability and responsibility can be shared fairly between all.

Arguing in favour of a closed-source licence also shows perhaps undue trust in private actors. It assumes that having control will mean that control will be used responsibly. A contrary argument would be fair, that is, that these private institutions do not warrant this trust. The fact remains that untrammelled access to such technology is a greater risk than a morally problematic private institution that maintains control and grants access to the system with some level of discrimination—a private agent (such as an institution or individual) that maintains exclusive control of the licence remains a better option than universal access, which would lead to an absolute guarantee of system (mis)use by malicious agents (such as the authorities of states with poor human rights records).

8.4 Fiduciary Theory, Responsibility, and Accountability

Accountability and responsibility have been discussed extensively in the preceding sections, therefore it is unnecessary to labour over these concepts too much once again. Here, accountability and responsibility will briefly be framed more specifically within the fiduciary context, before then proceeding to examine responsibility and accountability in the more functionally illustrative context of IHRL and its applications. This section will conclude by analysing the obligations of the fiduciary in its use of systems such as Slándáil vis-à-vis responsibility and accountability.

8.4.1 Responsibility, Accountability, and the Fiduciary Relationship

In the fiduciary relationship, as the fiduciary has been entrusted with power to rule on behalf of the public, it is thus accountable to the public for its use of power. More precisely, the fiduciary is "... accountable to public, fiduciary standards" (Criddle and Fox-Decent, 2012, p. 80). The use of the term "standards" is important here, as it establishes the normative basis of accountability, and signals that there are rules to which the fiduciary is held and can be challenged for breaking. If the rules are broken through the fiduciary's improper conduct, they are accountable to the public they are sworn to serve. The legitimate authority of the fiduciary is predicated on its provision of secure and equal freedom under the rule of law, with human rights serving as the blueprints to this regime, as has already been discussed on numerous occasions here—this is the object with which the subject has been entrusted. For the purposes of this section, human rights are argued to be constitutive of these fiduciary standards, and any failure to respect or enforce these rights represents a transgression for which they can be held to account either by the public, or, as secondary guarantors of human rights, international institutions such as the United Nations (Criddle and Fox-Decent, 2009, p. 385).

As an underlying concept of fiduciary authority is that an agent may not be judge and subject of the same cause (Fox-Decent, 2011), this also has numerous structural implications for the form accountability takes; it implies transparency so that the fiduciary's decisions and reasoning behind them are visible and might be challenged by the public, it implies a separation of powers so that the public can legally challenge the executive and legislative arms of the fiduciary, and it implies an independent police force so that any executive and legislative impropriety can be investigated and prosecuted[206]—the list in by way of example and not exhaustive.

Accountability is of fundamental importance in the fiduciary relationship, as the fiduciary's duty is to provide a regime of secure and equal freedom free of instrumentalisation and domination, sufficient safeguards, or mechanisms (and the basic conditions) of accountability also need to be in place in order to ensure that the fiduciary is sufficiently deterred and prevented from itself becoming a threat to the secure and equal freedom of its subjects that might instrumentalise or dominate them.

[206] And sufficient separation of departments within the police force in order that members of this institution can be investigated too.

8.4.2 Accountability and International Human Rights Law

Whilst the very existence of the international human rights regime contributes to a culture of accountability by providing normative rules or standards to which states should adhere (and are obliged to adhere to as a demand of their fiduciary role) and against which they can be challenged where they fail to do so, explicit obligations are outlined in many international human rights treaties and human rights practice in general that enshrine remedial and investigative obligations. In enshrining these investigative and remedial obligations, states are effectively required to investigate the causation of particular serious human rights abuses and prosecute responsible individuals, and provide remedy to victims of human rights abuse. Such legal obligations require states to ensure that there is effective national machinery in place domestically for victims of human rights, so that the state can be responsive to abuses and make necessary changes where required, and provide redress to those who have suffered from their failure to enforce and/or respect human rights, including taking "... appropriate measures toward the perpetrators [who are responsible for the human rights violation]", and ensuring that they are "...prosecuted and tried" (Joyner, 1997, p. 619).

The *UDHR* (1948) guarantees a right to effective remedy in Article 8. Additionally, the right to effective remedy is enshrined in Article 2, paragraph 3 of the *ICCPR* (1966). In the *ECHR* (1950), this right is enshrined in Article 13.

A practical requirement of the right to effective remedy is the duty to investigate violations of fundamental human rights (Joyner, 1997, p. 592; Mowbray, 2002; Van Dyke, 2005). Although no such obligation is explicitly stated within any texts known to the researcher, in practice, by courts such as the ECtHR, it is a noted procedural requirement.[207] This point will be revisited presently.

Examining the ECtHR's case law is once again instructive in understanding the practice of the right to effective remedy in an applied context. An important point to note from the start is that the right to effective remedy can be violated even where no other convention rights have been so. The Court merely requires that the claim of a human rights violation of any other convention right be an "arguable" one,[208] and as such even

[207] Though in the case of the *ECHR*, effective investigation is distinct but necessary element of effective remedy.

[208] According to Harris et al. (2009, p.561 citing *Boyle and Rice v. The United Kingdom,* [1988]):

if it does not find a violation of any other article on which the application was made, this does not preclude a violation of Article 13 if it is found that the responding state's legal machinery for providing remedy to the applicant was inadequate (Harris *et al.*, 2009, p. 560).[209]

In terms of Article 13's substantive requirements, persons must be able to bring forth their complaint of a human rights violation to a national authority which can offer remedy, which should be effective both in practice and in law (Harris *et al.*, 2009, p. 562). This remedy must prevent the human rights violation and/or its continuation, or provide redress to the victim (Harris *et al.*, 2009, p. 562). Where a remedy is provided "...by dint of the exercise of political discretion..." it "...will not suffice...", nor will advisory bodies that advise the final decision-makers in a given case constitute an effective remedy, as enforceability is a general requirement (Harris *et al.*, 2009, p. 563).

The above is not to imply that engagement with and decisions of authorities presenting possible avenues of effective remedy need to be judicial, though their power must be sufficient to guarantee a binding, effective remedy (Harris *et al.*, 2009, p. 565).

The institution acting as a potential avenue of effective remedy must also have sufficient independence from that which has been alleged to be responsible for the article violation in question (Mowbray, 2002, pp. 438–439; Harris *et al.*, 2009, p. 565). For instance, Harris *et al.* (2009. pp 565-566) cite *Khan v UK* [2000] as an example of a case where the institution was not sufficiently independent. Here, the applicant's options for complaints against the police were decided by the Court to not be sufficiently independent, as Harris *et al.* (2009, p. 566) explain that:

> ...on the facts the local Chief Constable had a discretion to refer matters to the Public Complaints Authority, failing which the standard procedure was to appoint a member of his own force to carry out the investigation. Further, as regards the Police Complaints Authority itself the Secretary of State had an important role in appointing, remunerating and, in certain circumstances,

---

> No abstract definition of the notion of arguability has been provided. The Court insists that, arguability 'must be determined, in the light of the particular facts and the nature of the legal issue or issues raised, whether each individual claim of violation forming the basis of a complaint under Article 13 was arguable, and, if so, whether the requirements of Article 13 were met in relation thereto'.

[209] For one example of this, see *Bubbins v. the United Kingdom* [2006]; here the Court heard a case regarding a police shooting the resulted in fatality, and while finding no violation of Article 2, did find a violation of Article 13 "..as the domestic legal regime was inadequate owing to lacunas in the compensatory regime" (Harris *et al.*, 2009, p. 560).

> dismissing its members, plus he had an influence on the withdrawal or referring of disciplinary charges and criminal proceedings.

Where an aggregation of possible avenues exist for potential effective remedy, the Court has found that the requirements of Article 13 are met; although, as pointed out by Harris *et al.* (2009, p. 567), in direct reference to *Leander v. Sweden* [1987], where several avenues are available but no individual one would seem to suffice:

> It is not made clear how each of the remedies reinforces any other. If any of them individually was adequate to satisfy Article 13, then no reference need be made to the others. On the other hand, if none of them individually were sufficient, as the dissenting judges thought, and none were appeals from another, then aggregating the series of inadequate measures would not be satisfactory to an applicant in the absence of an application of how the deficiencies of one were made up by the advantage of another, which the Court did not give.

Where national security is concerned, the Court has given states something of a greater margin of appreciation in matters of the right to effect remedy; for example in the case of secret surveillance it has found that remedies need only be accessible to an individual after the measures have been revealed, otherwise the access to remedy would undermine the measures taken to protect national security (which are of course legitimate where they meet the requirements, as discussed, of human rights limitations) (Harris *et al.*, 2009, p. 568).

Articles 2, 3, and 5 essentially have in-built requirements of effective remedy, though of most interest here is Article 2. Article 2 has been found by the Court to require effective investigation of violations of the right to life, and prosecution of responsible individuals—this procedural requirement essentially supersedes Article 13, which, broadly speaking, the Court may not examine once it has already examined the procedural elements of an Article 2 violation (Harris *et al.*, 2009, p. 573). With that being said, a failure to investigate undermines the right to remedy guaranteed by Article 13, therefore this will also be implicated in any ineffective discharge of the procedural requirements of other articles, so Article 13 in conjunction at least with Articles 2 and 3 do not necessarily function entirely independently of each other (Directorate General, Human Rights and Rule of Law and Council of Europe, 2013, p. 34).

Chapter 4 reviewed that Court's approach to the procedural requirements of Article 2 in some detail, pertinently, in the particular contexts of natural and man-made disaster. To remind the reader, in Chapter 4 it was established that the state has obligations to

safeguard human life, to the greatest reasonable extent within their capability, and to launch sufficiently independent investigations of the failures leading to loss of life (such as in natural or man-made disasters), investigations that can lead to prosecution of individuals found to be responsible where the conviction is significant enough to act as a deterrent from future such negligence or human rights violations. Notably, there is something of a threshold to negligence, with the Court deciding in *Jasinskis v. Latvia* [2011] that states have neglected their Article 2 obligations where, as noted by the Council of Europe (Council of Europe, 2013, p. 32 citing *Jasinkis v. Latvia,* [2011]):

> ...negligence attributable to State officials or bodies goes beyond an error of judgement or carelessness, in that the authorities in question, fully realising the likely consequences and disregarding the powers vested in them, have failed to take measures that have been necessary and sufficient to avert the risks to the victim's life.

An additional requirement of effective investigation requires that the authorities effectively secure evidence which can be used to establish causes of death and the person or persons whom might be responsible for said death (Mowbray, 2002, p. 439). Intuitively, where evidence collection is inadequate, an investigation may not lead to the identification and prosecution of responsible individuals, accountability is undermined and the prospect of redress for surviving relatives (required also to be involved in Article 2 related investigations, itself one of the demands of public scrutiny (Mowbray, 2002, p. 439)) will be dimmer (hence, Article 13 will be implicated).

In combination, the procedural requirements of Articles 2, 3, and 5 with Article 13 would seek to improve accountability by aiming to ensure that lines of causation of human rights abuse can be firmly established, responsible persons can be identified and subsequently punished, and redress is available for human rights abuse victims or their families. Where an adequate culture of accountability exists, this can bring about greater adherence to human rights through the inevitable requirements of reform where a state might identify problematic domestic legislation and general relations between its agents and subjects. Accountability importantly requires impartiality and independence (consistent with fiduciary characteristic that organs of the state cannot be judge and party to the same cause), and effective institutions that can secure positive outcomes for victims of human rights violations. The structure of human rights obligations, as decided by the practice and case law of the ECtHR based on the content of the *ECHR*, show how the fiduciary can be held to account in a manner that proscribes their use of power.

In what follows, the implications for effective remedy and the outlined procedural requirements of Article 2 of Slándáil-type systems will be revisited in light of the additional learnings made under Article 13, unexamined in Chapter 4.

8.4.3 Slándáil-type EMIS, Human Rights, Responsibility, and Accountability

To begin with, it is notable that the introduction of Slándáil-type systems to emergency management agencies will not alter the existing avenues accessible for effective remedy. These institutions will already exist, or not, and will already be effective, or not. A state which has few avenues available for effective remedy for any given human rights violation will obviously not gain any additional institutions with the introduction of Slándáil, however, it does have the capacity to make the work of existing institutions either easier or harder.

Slándáil-type systems and their supporting hardware and accompanying software can support investigative activities and avenues of effective remedy through the enhanced, digitally powered collection and preservation that modern technological solutions provide. Social media messages can be collected and stored as evidence of sources of intelligence that emergency managers can refer to in any public inquiry into decisions made during natural disaster response. In addition, where access to Slándáil-type systems is restricted to specified users and where user access is logged, either on the property of the emergency management agency where the system is completely housed on-site, or by remote system administrators and their systems where it is hosted remotely by creators/licensors, the persons responsible for use of the system and potentially responsible for taking action based on information received are logged and potentially accountable.

The preservation of social media messages collected by the system is important in order to help investigative authorities discern whether emergency managers were justified in making a given action based upon this information, and also whether the originator of the information (the social media user) should be brought into the investigation and potentially even prosecuted where it can be determined that such misleading information (if this is so) was posted with intent. With digital preservation of records, lines of causation become more clear than they might otherwise be. It is also important that social media messages be collected and stored in order for it to be transparent what the emergency manager knew, and if they could have prevented any tragedy that occurred based on what they knew (or were simply alerted of without taking any action

to verify the information). Access to the signals provided by Slándáil-type systems grants emergency managers potentially more knowledge of a situation than they might otherwise have, and they therefore become increasingly responsible. Proving negligence might still be difficult, and in many cases any action taken with negative outcomes may not even qualify, given the potential for large volumes of information (again, from multiple sources) being received by emergency managers combined with the limited resources at their disposal to react to any given reported incident.

Of course, for such systems to support effective investigation such digital solutions (and organisational solutions) need to be active. If user access to the system is not logged, and social media messages not stored, the situation may be more opaque. It will likely remain the fact that the emergency management organisation will have a list of persons who held access to the system, and will be aware of the working hours of those persons—such evidence may not be as clear and transparent, nor support fast and efficient investigation, as user access logging, but remains a viable solution. Furthermore, where social media messages are not collected emergency managers may struggle to provide a rationale for decisions made, and risk being held responsible for negligent action where they might otherwise have the possibility of arguing that they had a strong reason for making that decision (based on the content of a social media message that they saw but is later no longer available). Recording of the internal operations of any Slándáil-type system would also be important in order to potentially exonerate emergency managers of any harms arising from system failure, to the extent that it can be determined that said emergency manager was not responsible for or could not have averted the system failure.

Whilst the preceding was written with concern for investigations into violations of the right to life, once again, digital record keeping including what amounts to employee surveillance can serve investigations into other human rights violations. Consider the right to privacy, if it can be established that information obtained through Slándáil-type systems has been inappropriately shared, in a manner that illegitimately interferes with an individual's rights (that is, beyond what limitation or derogation might permit), by having in place mechanisms that can trace inappropriate disclosures, investigative authorities can identify responsible agents of the state. That is to say, if technology logs access to the Slándáil-system, perhaps even extending to monitoring personnel activity on the emergency management agency's systems and hardware, it would be possible to

identify individuals who were in contact with the information of an individual that may have been, by way of example, inappropriately disclosed.

Digital record keeping and employee surveillance would not be impervious to manipulation, of course, as records can be edited or deleted. Any such digital mechanisms put in place to support accountability through evidence collection would need to be constructed in such a manner that individuals directly responsible for human rights violation do not have unfettered access to the management of such records. Ideally, records and logs should be forwarded to a sufficiently independent body to maintain or inspect for a limited period of time.

There will be a limit to the efficacy or the potential contribution of any digital evidence trail towards assisting investigation or the realisation of effective remedy. Such evidence is only useful where there is a sufficiently independent and effective national authority available to receive the information and investigate it on behalf of persons claiming a human rights violation. Slándáil-type systems cannot compensate for a vacuum of national authorities designed to address and contribute towards the redress of human rights violations, it can only assist (and in conjunction with other technical and organisational mechanisms) by way of providing evidence for any such authorities to investigate or on which they may make a decision.

The creators/licensors also play a role to the extent that they provide a service utilised in a public context, and especially where they themselves collect social media data relating to a disaster, acquire the features of a public agency, at least by proxy, and should be bound by fiduciary standards, or at the very least held responsible for any actions that fail to respect human rights. For this reason, any level of employee surveillance applied to emergency managers should also be applied to actors within these private organisations with access to personal data. If data is disclosed or used inappropriately by these actors, in a manner at odds with national law, they too should be held accountable. If the system licensed is unfit for purpose based on negligent programming or design, and its use results in human rights violation, the appropriate persons within the organisation(s) should be identifiable and prosecuted where necessary, and the organisation or its employees/executives should not be able to evade liability from civil or criminal action if it can be established that actions taken by it played a substantial causative role in a human rights violation.

In view of the fact that such systems can be utilised to collect information about persons outside of the state experiencing a natural disaster or emergency, any party collecting and storing data should be open to communication from individuals outside of their national borders regarding information held about them, try within the best to their ability and to the extent that it is possible to identify information relating to that person, and anonymise or delete it if there is no justifiable reason for retaining it. The states engaged in such activities will need to grant persons outside of their borders access to institutions that can provide them with effective remedy where they have an arguable claim that their rights have been violated.

8.5 Conclusion

This chapter served the most critical role in the present research. Whereas previous chapters were concerned primarily with the boons and potential adverse consequences of Slándáil-type systems vis-à-vis the respective values used for analysis, here, by employing an accountability/responsibility based analysis, the discussion emphasised and examined how sources of evil might be identified, evaluated, and subsequently acted upon in order to punish and prevent evil.

This chapter also served an important role in disentangling accountability from responsibility, and by emphasising the differences between both in order to fairly counter the so-called challenges to accountability posed by modern information and computer systems. It was argued, following Floridi (2013) and to an extent Stahl (2006b), that AAs, and not just humans, could be accountable where conditions permit, but in such cases, responsibility dictates that the AA must be re-engineered such that it is no longer capable or pre-disposed towards harm.

In the preceding analysis, the importance of digital record keeping was emphasised as it serves as an evidential base that can serve either to exonerate or condemn agents. Clear lines of role responsibility, and roles, combined with digital records support accountability by making it easier to identify agents that have deviated from the norms of their roles, or broadly social/moral/legal (as in the more specific human rights case) norms.

The learnings here will make clear contributions to the guidelines that follow in the next chapter. It has become apparent, from the foregoing analysis, that any Slándáil-type system would be ethically most effective where roles, their duties and limits, are clearly

assigned to professionals involved in their design and use, and that technological solutions can be used to record the actions of agents (artificial and human) at all levels of the moral situation.

9 GUIDELINES FOR THE DESIGN AND USE OF SOCIAL MEDIA POWERED EMERGENCY MANAGEMENT INFORMATION SYSTEMS

9.1 Introduction

In this chapter guidelines for the development and deployment of a Slándáil-type EMIS that respects the dignity, moral values, and human rights of persons embedded in or otherwise affected by the MAS of which such systems constitute will be outlined. These guidelines are extrapolated from the discussion of the previous chapters as well as reaffirming important points already raised.

The guidelines presented here are not intended to be exhaustive nor monolithic. The research conducted, whilst extensive, was not exhaustive in itself with consideration for temporal and space limitations. Values were chosen based on perceived importance, and the guidelines that follow are based only upon the values analysed. Values may remain that warrant analysis. Values themselves are also not monolithic. Morals change throughout time and space. The task of this research was also predictive by design. Its goal was to pre-emptively establish the implications of Slándáil-type systems before general deployment, in order to predict potential threats to human rights and dignity and contribute towards discourse that could help mitigate such threats before they come to pass. Issues in system design and usage may well manifest that were not predicted here, with either positive or negative implications for the values chosen.[210]

These guidelines are also presented with some level of generality for the greater part, and they are not intended to be an instruction booklet demanding strict adherence, without which moral action cannot succeed. The guidelines are presented with a level of generality in order to support human autonomy, particularly amidst dynamic circumstances and in a world where technological and social landscapes are shifting quite quickly. The point is to guide the relevant agents in a direction that supports the moral design and deployment of such systems, accepting that not all variables can be anticipated and that these agents will know the circumstances of their environment, and

[210] With all of this in mind, the researcher asks that this be considered an organic list, one which invites debate, as well as further research on Slándáil-type systems as their impacts become empirically testable, both qualitatively, and quantitatively.

the resources at their disposal, better than the researcher. The researcher does not seek to unreasonably restrict their own capacity to determine the best solutions to moral problems, merely to provide a suggested path, based on rigorous study, down which they might travel with ethical goals in mind, aware of the boundaries of that path and that deviating too far from it may result in the agent getting morally lost.

As to whom these agents are that the following guidelines are intended to be read by, these are those involved in the creation and design of Slándáil-type systems, emergency managers, as well as the relevant legislators and executives whose role it is to shape the law governing such systems in a reasonable way, as well as ultimately declaring the states of emergency and subsequent derogations that they may require.

As a final note, some of the guidelines presented here were arrived at in research published within the first year of the Slándáil project, which can be read in the conference paper, *Ethical Framework for* a *Disaster Management Decision Support System Which Harvests Social Media Data on a Large Scale* by Damian Jackson, Carlo Aldrovandi, and the present researcher, Paul Hayes (2015). The curious reader is encouraged to read this work. The present guidelines, benefiting from being extrapolated from research conducted over a greater timeframe and with a different theoretical framework, expand and elaborate greatly upon areas analysed in previous work. In some cases, guidelines presented here will even contradict guidelines found in the *Ethical Framework*; where this is so, the contradictions shall be noted and explained.

9.2 Privacy

In Chapter 5 it was established that social media messages are rich in personal information, not just of the message originators' but potentially of third parties who are discussed or are present in images. It was established that due to deviations from context particular norms, the introduction of Slándáil-type systems could be classified as a *prima-facie* privacy violation, though on balance, with regard to the potential positive societal impacts entailed, their use could be justified. Nonetheless, potential for misuse remains, and the following points are made in an effort to mitigate indiscriminate or otherwise unjustifiable interferences with the right to privacy:

- Technological solutions should be applied to the greatest reasonable extent possible to protect the privacy of persons whose data is collected during the process of emergency management, without compromising the quality of

information obtained during natural disaster response. With regard to the importance of identifying information in emergency response (it is important to be able to identify people who may be in trouble), anonymisation may be too restrictive. The example of the Intrusion Index is indicative of approaches that may be taken.

- The infrastructure of such systems should support authorised user log-ins in order to limit the number of individuals with access to personal information. Only persons with legitimate reason to access such information, as per the requirements of their role, should be granted this access by end-user organisations, including emergency managers or investigators involved in any subsequent public inquiry. To the extent that system designers/licensors may also collect social media data during emergency, the number of persons with direct access to such information should be limited to people whose role necessitates it.
- Licensers of Slándáil-type systems should implement technical measures restricting the geographical search areas of such systems to the state borders of the licensee of the system, at least pending agreement by a second state to extend the search boundaries into their territory—and only in such a case where it is understood that the system will be used in the commission of cross-border natural disaster management.
- Any entity hosting personal data collected by the system should ensure that the latest most effective hardware and software solutions are utilised to secure the data from external intrusion, such as by hackers.
- Personal information obtained on social media should only be transmitted as necessary to achieve the goals of emergency management, that is, where information disclosure serves the purpose of saving life and property (for example, it may be necessary to share a photograph in order to locate missing persons).
- All entities hosting personal information should be responsive to requests by the public relating to data held about them. This includes requests from outside the state's borders. Information held relating to this data subject should be deleted at their request if retention serves no justifiable purpose—or appended with correction where it is justifiable but incorrect.

- Personal information should not be held longer than necessary to discharge the duties of emergency management, and any subsequent evaluation, public inquiry, or prosecution, which should be conducted without delay. Archived messages, including photographic or video content, should be anonymised if retained by any entity longer than necessary for the discharge of its duty or responsibility. This point is particularly applicable to designers/licence holders, who may retain social media messages in order to train systems. Training data should be anonymised.
- The system should only be activated by emergency management agencies where there is a clear threat emanating from natural hazards, and only for so long as that threat persists and poses a significant danger to the population. The geographical area of collected tweets should extend only to the area under threat. System use should be authorised by national law of a satisfactory quality, the implications of which are clear to social media users and persons whose data may otherwise be processed.
- Where there is doubt as to the quality of national law, the state must declare a state of emergency and notify relevant international bodies of the measures being taken impacting human rights, that is, which articles of relevant human rights treaties are being derogated from. The threat must be sufficiently grave, and the measures implemented only for so long as necessary. A declaration of emergency by emergency management agencies is insufficient—it must come either from the legislative, or executive, as applicable. Lawmakers are encouraged to craft law of sufficient quality to authorise use of such systems without derogation.[211]
- It is assumed that national law authorises retention of personal information for as long as necessary after the emergency has passed, as required for any investigative activities or public inquiries. Where this is not so, data should be deleted or anonymised as soon as possible.

9.3 Justice

Chapter 6, analysing justice with a focus on equality and discrimination, showed that vulnerable populations are more deeply impacted by disaster, and simultaneously, may

[211] Such law should not be written so as to authorise perpetual activation of such systems. Perpetual use would not be justifiable (it would be difficult to defend the proportionality of which), and would remain a human rights violation, not a justifiable interference.

not be present on social media, thereby biasing the information available to emergency managers and leading them to make decisions benefitting more resilient communities. It was established that depriving vulnerable groups of aid could be an act of either direct or indirect discrimination. The following points will propose suggestions in an effort to mitigate such problems:

- System designers, particularly of EMIS such as SIGE, should ensure that their system can support a diverse set of national data—that is, it should be able to load national data sets (including Census data), and display it as layers on its interactive maps. A degree of customisability may need to be provided to end-users, or the licensors should be responsive to requests to update data-sets.
- It is important that end-users be provided with a diverse range of data-sets that can be overlaid on interactive mapping functions, including information relating to environmental risk, important resources and infrastructure, and extensive population socio-economic data. End-users should ensure that they obtain this data even if it is not provided as part of the service. It is also important in order to avert biased emergency response that infrastructure information pertaining to mobile networks (such as mast locations and effective radii) is provided so that it can be determined if there are any areas that are unlikely to have access to the internet in order to use social media.
- Computational methods can be used to determine area risk by weighing different variables, as discussed in Chapter 4, including social media in the weighing. With the knowledge that certain vulnerable communities experience greater impacts from disaster, any such computational methods should pay particular attention to these communities when weighing risk. The Haase-Pratschke All Ireland Deprivation Index was provided as an example of a static statistical combination of variables associated with vulnerable communities that defines deprivation scores. Even where a computational method is not devised to automatically generate risk scores, static data-sets that can index the severity of socio-economic vulnerability in an area should be provided by licensors or otherwise obtained and utilised by end-users.
- Expanded language support should be an ongoing task by system creators. It is important that no voices are excluded on the basis that they do not speak the common tongue of the state experiencing an emergency. This might extend to automated translation support for emergency managers, who cannot be

expected to be extensively multilingual, at least beyond the prominently spoken languages of their jurisdiction.

- The most vulnerable populations, including but not limited to; older persons, low income households, women (in countries with sharp gender inequality), single-parent families, the disabled, renters, and persons living in inadequate housing units, should be the prioritised beneficiaries of emergency response to natural disaster (particularly where these categories overlap). Again, a Slándáil-type system should be able to convey to emergency managers where such population groups are located in high concentrations.[212]
- With this in mind, where data is not available to emergency managers pertaining to these population demographic categories, it is vital that needs based assessments sensitive to socio-economic vulnerability are conducted as a matter of emergency planning.
- Emergency managers should seek information from a plurality of sources, and not allow themselves to be lead by information obtained from social media, which may not be representative of all population needs.
- Emergency managers should make their populations aware of the offline capabilities of social media services (primarily Twitter) so that persons without internet access, or using obsolete technology, can engage with social media where emergency managers are utilising Slándáil-type systems.
- The system should not under any circumstances be used to interfere with the privacy rights of minority population groups, such as targeted monitoring of such groups that does not serve the goals of emergency management.

9.4 Trust

Chapter 7 of this research explored the trust implications of Slándáil-type systems with a particular interest in the problems of mis/disinformation and function creep. The guidelines presented in this section are suggested in an effort to mitigate the possibility of unethical consequences of such problems, and to mitigate their negative impact on trust between all agents, including artificial, involved in the moral situations explored here.

[212] This is in contrast to (Jackson, Aldrovandi and Hayes, 2015, p. 177) where it was argued that social sorting is unethical. On the contrary, with distinct qualification, Chapter 6 demonstrated that social sorting is acceptable, if not indeed ethically mandatory, where measures of social sorting are used to benefit society's most vulnerable.

- System creators should implement technological solutions within systems that harvest data from social media that can automatically gauge the credibility of processed information and/or the trustworthiness of originators of such information. Such automated systems should be sensitive to social media accounts that repetitively post false information, possibly flagging them for blacklisting as information sources as appropriate and to the extent that this is possible.
- It is desirable, given the potential of natural disasters to develop into technological disasters, that system creators provide dictionaries that enable systems to detect information on social media streams pertaining to technological hazards and other man-made disasters. The addition of functionality enabling it to detect information pertaining to criminal acts during disaster response, such as looting, is morally risky but justifiable.[213] Crime prevention, however, must be secondary to securing life in a disaster aftermath and any diversion of resources towards crime prevention such as preventing looting—particularly where it represents persons appropriating necessary supplies in times of dearth—and where this diversion of resources endangers persons in distress or in need of rescue, is morally wrong.
- Ideally, and in the interest of true democratic deliberation, the implementation of Slándáil-type systems will be put forth by the state for public consultation so that a plurality of actors can engage with and debate the relevant issues which concern the public interest.
- Any subsequent extension of functionality of the system, where such additional functionality is sufficiently removed from that necessary to achieve its original goals, should not occur without additional ethical and legal analysis, as well as additional public consultation.
- The system should, again, not be active outside of emergency or anticipated emergency, nor augmented to detect or suppress either criminal or legitimate

[213] In (Jackson, Aldrovandi and Hayes, 2015, pp. 177–178) it was highlighted, just as in here, that such use was essentially a slippery slope that could undermine public trust. However, as there remains a possibility of the breakdown of obedience to the law during times of disaster, and with knowledge that the State's duty is to provide a regime of secure and equal freedom under the rule of law to its subjects, and that regime can be challenged by malicious actors post-disaster, the State is justified in increasing the visibility of information pertaining to crimes that are an affront to its subjects' rights (at least insofar that it can demonstrate the under the exigencies of the situation, normal methods for achieving this are insufficient), particularly as the State is under positive obligation to protect these rights.

political activity during normal times. It should never be used under any circumstances to monitor and suppress legitimate political activity.

- Emergency managers should use all reasonable means at their disposal to verify information presented by Slándáil-type systems, regardless of whether or not it uses automated credibility analysis. It is important that emergency managers are confident in received intelligence before acting on it, or sharing it with the public.
- Emergency managers should address and correct false information in order to trigger a correction signal, where appropriate.

9.5 Responsibility and Accountability

Chapter 8 extensively examined issues as they relate to accountability and responsibility in the development and use of Slándáil-type systems; examining challenges ranging from many hands, and bugs, to opportunities including closed source licences and digital records as evidential bases. The following guidelines will make suggestions that attempt to mitigate the threats implied, and capitalise on the opportunities:

- Role responsibilities of persons involved in the development and eventual use of Slándáil-type systems should be clearly defined in order to improve the transparency of who was responsible for what action.
- Appropriate employee monitoring and disciplinary/reward mechanisms should be in place in both contexts in order to flag and censure unethical actions and praise/reward work well done, executed within ethical boundaries. Implementation would be at the discretion of relevant organisations, though a balance must be maintained between granting employees autonomy, and monitoring and restricting their activities—it would be unethical for their agency to be unreasonably restricted, or for them to work under unnecessarily authoritarian or oppressive conditions.
- Formal ethics and legal training should be provided to employees in both contexts in order to support the development of their ethical judgement.
- The system should be rigorously tested, using state of the art methods in line with current best practice, before being licensed. The system should not be licensed where it can be determined that critical errors are likely, or the system is unlikely to perform effectively.

- The system should log internal processes to enable easy bug identification. The creators should provide bug reporting mechanisms for end-users. The creators should respond to requests by end-users to provide fixes for any bugs detected.
- The system, and all supporting systems relevant to both licensors and end-users, should record actions taken, including user log-ins, who held access to personal information, what they did with it, and the particular actions made by emergency management agencies and the individuals responsible for those decisions.[214]
- The creators should, where possible, ensure that end-users have the capability of archiving processed social media messages in order to provide evidence as a rationale for their decisions in any public inquiry or investigation that follows disaster response. This may be needed to pursue prosecution of individuals responsible for morally bad outcomes in disaster response, whether they are state agents or civilians. Ideally, and to the extent that it is possible, the creators should also provide automated methods for data anonymisation or efficient deletion for when such archived content has served its purpose.
- Creators should use closed-source type licences that enable them to retain some control over system use to the extent that the technology is not easily and freely available. To that end, licensors should not license the technology to agencies they have reasonable cause to suspect will use the system unethically. Clear terms and conditions should be outlined in the licence demanding ethical and legal compliance, and also indicating the applicable laws where appropriate. This should allow the licensor to terminate a license where it can be determined that the licensee is abusing the system.

[214] In (Jackson, Aldrovandi and Hayes, 2015, p. 175) it was argued that:

> The incorporation of a journaling function that records management history as well as a journal of transactions, for the purposes of review, simulation and training is, ethically speaking, a double-edged sword. There is a risk that end-users' decision making could be influenced by the knowledge that every action taken on the system is recorded. Nevertheless, on the other side of the coin, the fact that there is a record of every action taken by each end user enables decisions to be retrospectively reviewed and evaluated should they be found to have been sub-optimal in the extant circumstances. Such a record also mitigates the risk of scapegoating in such circumstances.

On balance, such digital record keeping provides an important evidential base in any public inquiries that follow natural disaster, and can serve as a mechanism of accountability by making transparent who made what decision, and potentially why. Such digital record-keeping should also encourage agents to act within ethical and legal boundaries.

- Where the creator's technology has been widely cloned or duplicated as generic products that are freely available, maintaining strict licence control may no longer be an effective mechanism of preventing misuse of such technology, and the creator may be justified in gifting their knowledge to the public domain.
- Creators should not use licences to evade accountability.
- Again, emergency managers should utilise a plurality of information sources during emergency response, never relying excessively on Slándáil-type systems or allowing themselves to be epistemically enslaved.
- Where Slándáil-type systems prove ineffective or critically inefficient, and make no positive contribution to emergency response efforts, emergency managers should cease use until they can be appropriately re-engineered.
- State legislators and/or executive should ensure that sufficiently effective and independent national authorities are available to receive and adjudicate on human rights complaints by individuals who argue that Slándáil-type systems and their uses are implicated in negatively impacting their rights. Due to the transnational potential of rights violation, such authorities must be available to provide remedy to persons outside of the state's borders.

10 CONCLUSION

10.1 Introduction

In this concluding chapter, reflections on the preceding research and some thoughts on its broader implications will be offered.

This chapter will revisit the concept of Human Security, arguing that its importance to emergency management is paramount and that viewing such a process through the lens of traditional security is ineffective—disaster management in particular is a multi-agency effort requiring enhancement of human capability and community resilience. This chapter will briefly explore how Human Security is supported by Information Ethics and Fiduciary Theory.

It will reflect on the contribution of the dual framework to this research, arguing that it effectively provided a persuasive lens through which to analyse the chosen values and issues. The framework was instrumental in the development of the guidelines as read in the previous chapter.

This chapter will reflect on the responsible design and use of Slándáil-type systems, arguing that whilst they stand to be of great benefit to emergency managers, they demand careful design and deployment as their risks to human values and rights remain great. It will be argued that adherence to the guidelines proposed in Chapter 9 may serve to mitigate these risks.

Finally, the limitations of the current research will be assessed as well as future opportunities.

10.2 Human Security, National Security, Natural Disasters, and Social Media Powered EMIS

Chapter 1 introduced the very important concept of Human Security. Whilst explicit references to the concept since have been nil, its meaning and goals have permeated through most subsequent chapters. Though this may not be immediately apparent, the following will outline how Human Security was a silent conceptual strand running throughout this research, as well as how the theoretical framework used supports its

goals and legitimises it, as well as its implications for the use of social media powered EMIS, particularly in natural disaster response.

Human Security and its goals were broadly defined in Chapter 1 as being freedom from fear and want. It represents a reconceptualisation of security as normally applied by the state, something which typically entails an emphasis on Schmittian protection of the state and its apparatus against external, violent, threats (or internal, as the case may be)—the existential dangers faced by the state as an entity demanding continuity are no longer the primary referent of security (though they remain no less important), and instead individuals and their flourishing are the primary referent (Bacon, 2016; Cameron, 2016; Popovski, 2016; Zack, 2016). Human Security is multifaceted and holistic; people, as referents of security, must have their needs met and must flourish through positive and negative freedoms; the demands of their agency and dignity must be met (Bacon, 2016; Cameron, 2016). People must not go hungry, uneducated, live in environments of political oppression, or vulnerable to the impacts of natural disasters which are themselves a profound danger to Human Security, a fact which should have been well conveyed in Chapter 6. The concept owes much to the work of Sen and Nussbaum with its focus on capability.

Human Security is not simply about official mobilisation against threats to state continuity, it is about meaningful development and requisite freedoms to live a worthwhile life of dignity, and in its holistic view of security it recognises that protection from natural disasters is an important yet complex project, requiring that people have adequate capabilities (Bacon, 2016; Cameron, 2016; Zack, 2016). The concepts of homeland security, or national security, which are a dominant paradigm in natural disaster management, are ineffective and inadequate at truly reducing population vulnerability to natural disasters.[215]

[215] As argued by Zack (2016, pp. 58-59):

> There is a big difference between those Human Security programs for disaster and development in vulnerable populations, and Human Security programs that are part of the sovereign government's national security apparatus... Both Human Security and Homeland Security share conceptual division between planning and response. However, they differ conceptually in the meaning of "security." Security can mean *civilian safety*, which on the Human Security model refers to the decreasing the disaster-related risk of already vulnerable populations. Or security can refer to a process of overt or covert police, military, and other government planning and response as protection against intended harm from some human beings—national enemies, criminals exploiting disaster and traitors—to the collective wellbeing or to the "homeland." This second

In Human Security, it is not simply the state's borders that require securing, it is the individuals and their well-being within those borders that must be secured (and as it is a universal concept demanding global action, individuals behind all borders through international co-operation) (Popovski, 2016, p. 95). It is apparent that any disaster management framework based on a technocratic approach to securing infrastructure and minimising casualties without paying heed to the greater requirements of human development, while it might have tactical successes, is strategically doomed in its failure to address the structural inequalities that lead to some being more vulnerable to disaster than others. Human Security is a multi-agency effort, and not exclusively the domain of those more directly involved in emergency management.

As a moral value and human rights analysis of the implications of Slándáil-type systems, for use in natural disaster management contexts, this research was guided by an interest in the protection of human dignity. The purpose of the research was to uncover the morality and human rights implications of such systems with a view to understanding their prospective benefits to humankind, and how any threats they might pose to human dignity themselves could be mitigated. And yet the analysis conducted was not done from a technocratic or militarised perspective. It was not a cost-benefit analysis concerned solely with potential added value of such systems to disaster management, to the potential efficiencies added to the work of emergency management agencies from a human resources or financial perspective, nor was it a quantitative analysis of their potential to save life and mitigate disaster impacts and losses. The human being, their dignity, and all that entails was the centre of the analysis throughout. This research did not proceed with simple national security in mind—its approach was more holistic and, again, human centred.

The dual theoretical framework chosen to centre the analysis of this research, consisting of Information Ethics and Fiduciary Theory, broadly supported the concept of and outcomes desired by Human Security and ultimately provided a moral and perhaps legal

---

sense of security is the meaning implied by the name of the US agency, the Department of Homeland Security, and it is rests on the same traditional understanding of security against which the Human Security Paradigm was founded. However, we should note that while this meaning of "security" the following aspects of security, neither does it provide them: a focus on development within vulnerable communities, an increase in political representation for residents of such communities, priority assigned to disaster preparation in vulnerable communities. Still, Homeland Security represents a form of security for the state and its interests, rather than for people who live in that state.

basis for the Human Security approach, with obvious particular regards for natural disaster management and the deployment of social media powered EMIS within.

Information Ethics, with its, according to Floridi (2013), ecumenical or holistic concerns is concerned with the development of the agency of human beings, their flourishing, as well as the flourishing of their environment (fundamentally, the flourishing of all existence). At deeper analysis (using the broadly compatible and supporting capability theory), positive and negative freedoms (and ultimately functionings that support Nussbaum's basic capabilities), are intrinsic to the satisfactory and fulfilling development and flourishing of the human being (Carter, 2016). This combination of theories entailed active measures and programmes of support for human development, upon which dignity is contingent, as well as negative duties of interference that would provide obstacles to such development. In this ethical framework, there was a symmetry with the demands of Human Security, and it supported an ethical analysis that could help ensure outcomes compatible with Human Security. Beyond this, it also provided moral justification for Human Security. And "justification" may be too light a term. If supporting human capability, as well as opposing any obstacles (or sources of entropy) to the development of human capability is not just morally good, but morally required, then the Human Security approach is not merely justified, but is a moral imperative for national and international communities who must support the active development and reasonable freedom of all people, and thereby as a correlative, ensure that no international or state level policies threaten this.

Fiduciary theory is a non-positivist legal theory arguing that the state's purpose, that which shapes its duties towards its subjects, is the provision of a regime of secure and equal freedom under the rule of law, with human rights as constitutive of this regime. It is the antithesis of the more positivist, realist Schmittian views—it does not rule with an iron fist, nor with its own survival and continuity as a matter of importance exceeding any other.[216] Human rights and security are distinct, and yet inseparable.[217]

[216] Though obviously it does remain important, for there to be a regime of secure and equal freedom the State most survive and secure itself to provide it, just not at the cost of gross human rights abuse.

[217] According to Popovski (2016, p. 96):

> Human rights and Human Security ensure both 'freedom from fear' and 'freedom from want' and allow people to live in dignity and safety... People need protection, not only of their lives and freedoms, but also of their well-being, property, employment, family, health, environment and so on—Human Security is achieved through the promotion and

Whilst distinct concepts, both Human Security and human rights share the same goals (the protection of human dignity and the advancement of their agency and well-being), and are mutually reinforcing, they are so intertwined that a tentative argument can be made that if the state is duty bound to provide a regime of secure and equal freedom under the rule of law, with human rights as the blueprints of this regime, then (by virtue of the near symmetrical overlap between Human Security and human rights), the state is duty bound to provide a regime consistent with the demands of Human Security to all its legal subjects. This argument is dense and cannot be defended in full here—it is beyond the scope of this research—however it remains a persuasive indicator that Human Security is a legal entitlement, constitutive of the state's duties and its relationship with its subjects.

With this in mind, there is a moral and legal (at least insofar as human rights and Human Security share significant overlap) basis for framing natural disaster management as a matter of national (and international) Human Security, security which acts for and not simply upon human beings, security that supports their growth and freedom and does not exist merely to protect them as faceless subjects of a supreme authority from external threat. This framing is important, as it asserts that natural disaster management is a complex process requiring building resilient communities, removing sources of inequality, and not merely responding to humans as infrastructure that needs to be safeguarded.

The Human Security approach also helps ensure that measures put in place during natural disaster response do not in themselves harm the very people they are supposed to benefit, and are implemented in somewhat of an equitable fashion. Human Security respects human rights, and as such demands appropriate respect for civil and political freedoms. Here, an extensive review of implications of social media powered EMIS was conducted under numerous categories (or values), through both an ethical and human rights lens. It was established that systems such as Slándáil could be misused to the extent that they represent a threat to persons' dignity and rights, through privacy violation, to assist in inequitable response, and more. The emergency manager operating under a framework compatible with Human Security must mitigate the

---

> realization of various kinds of rights.... Often it is the state's failure to ensure Human Security that violates human rights and vice versa.

potential of such systems to do harm to those they are charged with protecting, lest they themselves represent a threat to Human Security.

The work here endeavoured to identify threats arising from the use of Slándáil-type systems, and consistent with Human Security, provided in Chapter 9 guidelines with the intention of mitigating any harm they could cause, by specifying restrictions on use and suggestions for design of the system. Any introduction of new technology with ethical and human rights implications should follow such processes; their risks need to be understood, particularly before deployment, so that Human Security is not threatened. A traditional security approach may overlook these reflective processes, ignoring the importance of dignity and freedom in the mechanical interest of preserving order at any cost.

As a final point, Human Security requires that one be reflective of the role something such as Slándáil plays in disaster management. It is not a panacea. Its potential to reduce impacts of disasters is limited. It cannot substitute for poor disaster planning, and certainly does not compensate for the state failing to invest in the resilience of vulnerable communities. Human Security demands recognition that these systems would be but one small element of natural disaster management, and perhaps that planning and preparation remain the most important.

10.3 Assessment of the Dual Framework

The preceding analysis was conducted with the aid of two theoretical frameworks, as described in Chapter 2. This was a challenging approach to take, requiring the researcher to regularly cross the divide between ethical and legal theory as well as requiring the consideration of a wider range of issues (and academic sources) than would otherwise be necessary. The theoretical framework was a risky proposition, though a decision which in the end bore fruit.

Information Ethics provided an opportunity for exploration of the values chosen for disclosive analysis from a unique informational perspective. This macroethical theory served the research well in casting light on the importance of the contribution of artificial entities to the well-being of multi-agent systems, and the importance of designing artificial entities that improve both these multi-agent systems and the infosphere they constitute, without damaging them both. This was achieved using the idea of distributed morality and moral thresholds. In this research, concerned with

relations between distant agents, as mediated by and with outcomes potentially contingent on artificial agents such as Slándáil, the selection of IE directed an understanding that each player in a moral situation, human or artificial, plays an important role in its outcome and its actions need to be responsible and concerned with the health of the infosphere, including the well-being of the moral patients that inhabit it.

IE proved to be a useful foundation for examining the nature and relevance of the values analysed in this research, allowing for an elaboration of what they represent, what they mean to humans, as well as ultimately how the interactions of agents in a Slándáil-type system mediated moral situation might support or undermine such values.

IE did not bear the weight of the preceding ethical analysis by itself. Use of the theory is difficult because it remains in development, inviting debate and discussion on its future and application. At times, addressing the implications of Slándáil-type systems for moral values using IE was a difficult task, and answers were unclear. Solutions in such cases were achieved through supplementing supporting theory to fill in gaps left by IE. Numerous theories were utilised where they adequately supported IE. Here, Nissenbaum's CI gave additional shape to analysis vis-à-vis privacy, and Capability Theory and Prioritarianism supplemented an analysis under justice. IE was found in these cases to engage well with other theories, justifying their own content under a macroethical foundation and providing more normative form to IE (and supporting more rigorous and explicit conclusions). The minimalistic content of IE supports such engagement with compatible theory, and should be viewed as one of its strengths, a strength which in turn can support its own development or applicability looking to the future.

To the best of the researcher's knowledge at time of writing, the research here represents one of the more extensive applications of IE to a single case (the use of social media powered EMIS). It is the researcher's hope that this research has proven IE's value, its versatility, and its applicability. In utilising this theory, it has contributed to academic knowledge in a original way both by testing the theory through application, and by applying it to a very novel and developing case.

Whilst the underlying concepts of Fiduciary Theory are not new, the manner in which it has been applied to state authority by Fox-Decent and Criddle—a manner that argues

that human rights are an intrinsic aspect of state duty towards its subjects—is quite recent itself. It proved its use in the context of this research by explaining the sources of state authority, as well as, particularly within the context of times of emergency, describing and proscribing the limits of this authority, without falling into a Schmittian trap suggesting that there are no limits. The usefulness of this theory was in providing a theoretical foundation for the understanding and application of human rights, something which was important in directing consistent and effective analysis in a world with fractured human rights practice (consider the number of institutions worldwide that adjudicate or advocate on human rights). It provided a theoretical structure to propose clear solutions to human rights issues where ambiguity may otherwise obfuscate analysis, or weaken any arguments made. One point where the use of Fiduciary Theory shined in particular was in supporting a solution to the problem of applying human rights extra-territorially, a contentious topic, and one where its principles indicated could be solved by adopting the gestalt approach, that is, a state's human rights responsibilities are commensurate with the degree of control it holds over those subject to its power.

Fiduciary Theory was selected in combination with IE in order to, essentially, put a double-lock on the conclusions reached in the disclosive analysis. The research was concerned with pursuing the preservation of human dignity, particularly in times of disaster where it comes under such strain, from an ethical perspective. However reaching ethical justifications, the researcher realised, may not be enough to appeal to all potential actors (particularly within government agencies). It can be easy to dismiss the notion of ethics or moral theory; an actor under the pressure of exigent circumstances will likely always distinguish between what they ought to do, and what they are authorised to do within the parameters of law. Law and ethics differ. Ethics may, and often does not, enjoy the force of law. One can explain to a state actor what is right and just, but they may still retain wide discretion in what actions they execute so long as those actions do not fall afoul of the requirements of written law or any kind of executive or military order. In states around the world one may never struggle too much in finding examples of state actors committing heinous acts that are not ethically justifiable, but nonetheless ostensibly fall within some weak veil of legitimacy as they were authorised by the state's legislative or executive. One need only point to ongoing civil conflicts such as within Syria; or for a more historical example the "legally" authorised actions of the Nazis during the Holocaust.

To argue persuasive conclusions in such a context it can be useful to reach beyond ethics, or to root ethical conclusions more firmly in the principles that legally constitute a state's duties. Fiduciary Theory presented the opportunity to secure conclusions that supported some of the ethical analysis, insofar as it applied to state action, that may be more persuasive to state actors as they would be grounded in legal theory that more firmly indicated that some of these ethical conclusions were not merely based on meditations on what one ought to do, or should do, for their actions to remain ethical, but on what they *must* do so that they fall within the parameters of what they are legally authorised to do, based on the state duty to provide a regime of secure and equal freedom that fundamentally requires the respect and enforcement of human rights.

As with IE, the researcher is not aware of Fiduciary Theory being applied so extensively to a single case, and is absolutely positive that it has not been applied to this specific case. In utilising this theory, value was added to the research, granting the theoretical approach and conclusions novelty, and in so doing, rendering this an original contribution to academic knowledge.

The researcher hoped that there would be more opportunities for engagement, or dialogue, between theories—that IE and its ontology of information could supply Fiduciary Theory with additional moral authority and understanding of threats to human dignity. This was not comprehensively the case, though in Chapter 5, dealing with privacy, it was understood that violations of privacy, by virtue of the understanding that we are our information, could be instrumentalisation as understood by Fiduciary Theory and thereby underlined the moral importance of the protection of this right. Ultimately, both of theories served not to contradict each-other upon application, and were largely complementary in their understanding of the importance of human dignity and agency. This complementarity aided a coherent analysis of values.

10.4 Reflections on Social Media Powered EMIS: Deploying Them Ethically and in a Manner that Respects Human Rights

The preceding research emphasised the importance of Slándáil-type systems in their ability to bridge power and knowledge gaps between emergency managers and social media users, and how the power of distributed morality could be harnessed to potentially escape instances of the tragedy of the Good Will. With climate change rendering the threat of natural disaster ever greater, we as human beings—morally

responsible agents—need to deploy whatever ethical means we have at our disposal to mitigate its impacts. Systems such as Slándáil represent another tool in our arsenal to achieve this, and so long as they can be deployed ethically, they represent a moral imperative. They can potentially empower the otherwise powerless, and they can tap into the raw potential of data-rich social media streams. They represent a capitalisation of technology towards morally good, beneficent ends. Importantly, with their ability to record and preserve actions through journaling and bug logging systems, they can also support values of responsibility and accountability wherever negligent action arises when agents fail to appropriately discharge their duties.

The outlook is of course not entirely optimistic. Such technologies can be deployed towards unethical ends, or otherwise be used to threaten human rights or Human Security more broadly. Malicious actors can use such technologies to terrorise the innocent; to indiscriminately violate privacy, to target persons based on personal characteristics or help the privileged at the expense of the vulnerable—this list is not exhaustive. Even where evil is not executed with intent, such systems may simply fail to be effective with respect to either technical errors or the potentially insurmountable challenge of pervasive unreliable information present on social media feeds. The value threats should not be understated. Such systems can be used in potentially terrible ways, and society as a whole needs to be vigilant and hold to account those who design and use them.

In the disclosive analysis, using the dual theoretical framework, many value threats were explored and based on this exploration, guidelines—as seen in Chapter 9—were formulated in an effort to guide the value-sensitive design and deployment of Slándáil-type systems. Adherence to these guidelines by relevant actors should go some way to ensuring that such systems are designed and deployed in a manner that respects the analysed values. With that being said, such work still only represents words on paper, and can be dismissed by actors seeking to do harm, or who simply do not care that their actions cause harm, irrespective of the consequences for them. Guidelines such as these should be presented to as wide an audience as possible, for then the social media using public, and the disaster affected, can be aware of their rights and hold the concerned private and public actors to account. With this in mind, in some ways the work of this research has just begun, and for it to be truly fruitful it will need to continue, and to be

designed in a way that is understandable to yet a wider audience outside of an academic context.

In the most optimistic scenario, systems such as Slándáil will be designed ethically, with integrated tools that assist responsibility and accountability such as the Intrusion Index and a journaling system. They will function as effectively as possible. Importantly, they will only be licensed to trustworthy institutions and not the apparatus of regimes that hold little regard for human rights or the rule of law. They will be deployed only when a threat to a community is real and imminent, and only insofar as they are necessary to address the exigencies of that threat and only for as long as necessary to do so. They will not be active in perpetuity, nor will personal data be stored indefinitely. They will be deployed for the benefit of societies' most vulnerable, and not merely those who are already privileged or resilient to disaster. They will not evolve past their initially intended goals without adequate additional ethical and human rights analysis and public consultation.

The guidelines presented here, if followed, should allow emergency managers to reap the rewards of effective emergency management, potentially saving lives and important infrastructure by capitalising on the knowledge of the social media user, and doing so without threatening important values that ensure societies' flourishing, and peoples' agency and dignity. It is the hope of the researcher that the warnings offered here are headed, and that satisfactory efforts are made to prevent the widespread misuse of such systems as Slándaíl.

10.5 Limitations and Looking to the Future of Research on Social Media Powered EMIS

The research was conducted before the official deployment of the Slándáil EMIS in natural disaster. In the future, it may have been possible to assess the system under study post-deployment in order to witness, record, and evaluate how emergency managers were using such systems in practice. This would hold value, as subsequent research could recommend corrective action where-ever morally dubious, and importantly, perhaps unanticipated, system uses or related actions were being executed. This research however was predictive in nature, and served an important purpose in examining potential adverse value implications ahead of deployment in order to pre-emptively flag and potentially prevent any actions that would affect human

dignity or adversely impact human rights. In some ways, this is more important than retrospective analysis, although it does not undermine the necessity of both. Further research is invited, and will also serve an important role. Systems such as Slándáil should be under constant scrutiny. They have real impacts on human values, and should not be forgotten or ignored. Ignorance can only lead to misuse upon a complacent public.

Not all potentially relevant values were explicitly analysed here. A theme running throughout was that of transparency, and its omission here (due to limitations of time and space available) in its own chapter is unfortunate. This value is ripe for further research in this context, and is invited by the researcher. In the interim, the researcher is confident that the disclosive analysis is comprehensive enough to have supplied comprehensive guidelines, adherence to all of which should still facilitate ethical design and use of such systems.

Additionally, it has been argued that such guidelines are organic and subject to change perhaps throughout time and space. The researcher invites debate, if not contradiction, acknowledging that perhaps at times his conclusions were either too strict or lenient. Debate is welcome, the researcher hopes that this research may even serve as the bedrock of such debate, and hopes to engage with it, whether to concede points or vehemently defend them, as the case may be.

This research also highlighted that research of the theory used should continue, with particular regard for Human Security. It has been indicated here that IE and Fiduciary Theory provide a supportive, moral and legal basis for Human Security. It was not within the scope of this research to dwell on this, it was merely a useful and insightful by-product, and perhaps the start of something else to be continued elsewhere. The concept of Human Security should be engaged with by both moral and legal philosophers, it is a ripe area of study, particularly in these regards.

10.6 Final Thoughts

Ultimately, the preceding research demonstrates that social media powered EMIS have the innate potential for good in the context of the response phase of natural disaster management, and perhaps emergency management more broadly. They can harness the power of the crowd, of the observers on the ground who live through the impacts of destructive natural forces. Those who may ordinarily be powerless to individually effect change can use their online voices to help increase the situational awareness of

emergency managers who can in turn act on this information and use their power to mitigate forces of natural evil. Those with knowledge but without power can effectively join forces with those with power but not necessarily knowledge, through the power of ICTs and the hyperhistorical information life-cycle, to harness the power of distributed morality.

Such systems also support statutory agencies in carrying out their fiduciary duty of protecting life whilst also—because of their democratising nature—help the very public that is in danger contribute towards life saving actions, by volunteering information and enhancing situational awareness.

The preceding research also demonstrates that optimism should be tempered; this theoretical innate potential for good may not translate into reality. Adverse value impacts were implicated. Such systems can be complicit in violations of privacy, they can be used as cost-effective methods that allow statutory agencies to insert themselves into the online space towards unethical ends, they can cause the deterioration of trust in multi-agent systems, they pose challenges to accountability and responsibility, and they may exclude or even contribute towards the persecution of society's marginalised. Essentially, when misused, they can be used to attack dignity, not protect it.

The purpose of this research was to cast a light on the possibilities of social media powered EMIS, both good and bad. Only through interrogating these possibilities can we identify where things may go wrong, by identifying the worst case scenarios we can attempt to propose solutions that help mitigate them before they materialise. This research attempted that. It has been research that is optimistic about future applications of digital technologies towards life-saving ends, but has warned that caution in their development and deployment is necessary. We must always be cognisant of how such technologies are used, and always try to hold account those who design and use them. At worst, if uses of such systems have adverse value impacts that go unchallenged, such uses will be normalised and those who use their power towards evil ends will do so with impunity. It is our collective responsibility to challenge those in positions of responsibility themselves to operate within ethical and legal boundaries. We must hold them to account, and keep them honest. Accountability fails when the public itself is acquiescent or complacent. The results of this research are something for everyone to think about, not just software developers and emergency managers, but

the average citizen as well, whose rights are implicated in the use of social media powered EMIS.

BIBLIOGRAPHY

van Aalst, K., M. (2006) 'The impacts of climate change on the risk of natural disasters', *Disasters*, 30(1), pp. 5–18.

Adam, A. (2005) 'Delegating and distributing morality: Can we inscribe privacy protection in a machine?', *Ethics and Information Technology*, 7, p. pp.233-242.

Adam, A. (2005) *Gender, Ethics and Information Technology*. 2005 edition. Houndmills, Basingstoke ; New York: Palgrave Macmillan.

Agamben, G. (2005) *State of Exception*. Translated by K. Attell. Chicago: University of Chicago Press.

Aizu, I. (2011) 'The role of ICT during the disaster – A story of how Internet and other information and communication services could or could not help relief operations at the Great East Japan Earthquake'. Available at: http://www.ispp.jp/ispp-wp/wp-content/uploads/2011/09/EarthquakeICT0825.pdf (Accessed: 2 August 2017).

American Red Cross (2010a) 'Social Media in Disasters and Emergencies'. Available at: https://www.slideshare.net/wharman/social-media-in-disasters-and-emergencies-aug-5 (Accessed: 2 August 2017).

American Red Cross (2010b) 'White Paper: The Case for Integrating Crisis Response With Social Media'. Available at: https://www.scribd.com/doc/35737608/White-Paper-The-Case-for-Integrating-Crisis-Response-With-Social-Media#download (Accessed: 6 December 2016).

Ashktorab, Z. *et al.* (2014) 'Tweedr: Mining Twitter to Inform Disaster Response', in *Proceedings of the 11th International ISCRAM Conference*. *ISCRAM*, University Park, Pennsylvania, USA. Available at: http://cs.iit.edu/~culotta/pubs/ashktorab14tweedr.pdf.

Auf Der Heide, E. (2004) 'Common Misconceptions about Disasters: Panic, the "Disaster Syndrome," and Looting', in O'Leary, M. (ed.) *The First 72 Hours: A Community Approach to Disaster Preparedness*. New York: iUniverse, Inc., pp. 340–380.

Backman, C. (2012) 'Mandatory Criminal Record Checks in Sweden: Scandals and Function Creep', *Surveillance & Society*, 10(3/4), pp. 276–291.

Bacon, P. (2016) 'Incorporating natural disasters into the human security agenda', in Hobson, C., Bacon, P., and Cameron, R. (eds) *Human Security and Natural Disasters*. 1 edition. Place of publication not identified: Routledge, pp. 1–21.

Baier, A. (1986) 'Trust and Antitrust', *Ethics*, 96(2), pp. 231–260.

Balkin, J. (2004) 'Digital Speech and Democratic Culture: A Theory of Freedom of Expression for the Information Society', *Faculty Scholarship Series*. Available at: http://digitalcommons.law.yale.edu/fss_papers/240.

Ball, J. (2013) ‘NSA’s Prism surveillance program: how it works and what it can do’, *The Guardian*, 8 June. Available at: https://www.theguardian.com/world/2013/jun/08/nsa-prism-server-collection-facebook-google (Accessed: 5 February 2017).

BBC (2014) ‘Sina Weibo: “China”s Twitter’ to list in the US’, *BBC News*, 14 March. Available at: http://www.bbc.com/news/business-26588397 (Accessed: 2 August 2017).

Beran, H. (1977) ‘In Defense of the Consent Theory of Political Obligation and Authority’, *Ethics*, 87(3), pp. 260–271. doi: 10.1086/292040.

Bhattacharya, S. (2004) *Killer convicted thanks to relative’s DNA*, *New Scientist*. Available at: https://www.newscientist.com/article/dn4908-killer-convicted-thanks-to-relatives-dna/ (Accessed: 12 June 2017).

Birsch, D. (2004) ‘Moral Responsibility for Harm Caused by Computer System Failures’, *Ethics and Information Technology*, 6(4), pp. 233–245. doi: 10.1007/s10676-005-5609-5.

Brey, P. (2000) ‘Method in computer ethics: Towards a multi-level interdisciplinary approach’, *Ethics and Information Technology*, 2(2), pp. 125–129. doi: 10.1023/A:1010076000182.

Brey, P. (2008) ‘Do we have moral duties towards information objects?’, *Ethics and Information Technology*, 10(2–3), pp. 109–114. doi: 10.1007/s10676-008-9170-x.

Brey, P. (2010) ‘Values in technology and disclosive computer ethics’, in Floridi, L. (ed.) *The Camridge Handbook of Information and Computer Ethics*. Paperback. New York: Cambridge University Press, pp. 41–58.

Broadband Commission Working Group on Broadband and Gender (2013) *Doubling digital opportunities: Enhancing the inclusion of women and girls in the information society*. Geneva, Switzerland. Available at: http://www.unwomen.org/en/docs/2013/9/doubling-digital-opportunities-women-and-girls-in-it (Accessed: 6 April 2017).

Bynum, T. W. (2006) ‘Flourishing Ethics’, *Ethics and Information Technology*, 8(4), pp. 157–173. doi: 10.1007/s10676-006-9107-1.

Cameron, R. (2016) ‘A more “human” human security: the importance of existential security in resilient communities’, in Hobson, C., Bacon, P., and Cameron, R. (eds) *Human Security and Natural Disasters*. 1 edition. Place of publication not identified: Routledge, pp. 158–180.

Carter, I. (2016) ‘Positive and Negative Liberty’, in Zalta, E. N. (ed.) *The Stanford Encyclopedia of Philosophy*. Fall 2016. Metaphysics Research Lab, Stanford University. Available at: https://plato.stanford.edu/archives/fall2016/entries/liberty-positive-negative/.

Cassinelli, C. W. (1959) ‘The “Consent” of the Governed’, *The Western Political Quarterly*, 12(2), pp. 391–409. doi: 10.2307/443978.

Castillo, C., Mendoza, M. and Poblete, B. (2011) 'Information Credibility on Twitter', in *Proceedings of the 20th International Conference on World Wide Web*. New York, NY, USA: ACM (WWW '11), pp. 675–684. doi: 10.1145/1963405.1963500.

Chopra, S. and Dexter, S. (2009) 'The freedoms of software and its ethical uses', *Ethics and Information Technology*, 11(4), p. 287. doi: 10.1007/s10676-009-9191-0.

Corbet, R. *et al.* (2017) 'D2.6 Licence for the Use of a Disaster Management System'.

Council of Europe Research Division (2015) *Internet: Case-law of the European Court of Human Rights*, *Refworld*. Available at: http://www.echr.coe.int/Documents/Research_report_internet_ENG.pdf (Accessed: 7 June 2017).

Criddle, E. J. (2014) 'A Sacred Trust of Civilization: Fiduciary Foundations of International Law'. Available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2336075.

Criddle, E. J. and Fox-Decent, E. (2009) 'A Fiduciary Theory of Jus Cogens', *The Yale Journal of International Law*, 34, pp. 331–387.

Criddle, E. J. and Fox-Decent, E. (2012) 'Human Rights, Emergencies, and the Rule of Law', *Human Rights Quarterly*, 34(1), pp. 39–87.

Currion, P., Silva, C. de and Van de Walle, B. (2007) 'Open Source Software for Disaster Management', *Commun. ACM*, 50(3), pp. 61–65. doi: 10.1145/1226736.1226768.

Cutter, S. L., Boruff, B. and Shirley, W. L. (2001) *Indicators of Social Vulnerability to Hazards*. Columbia, S.C: University of South Carolina, Hazards Research Lab.

Dahl, J. Y. and Sætnan, A. R. (2009) '"It all happened so slowly" – On controlling function creep in forensic DNA databases', *International Journal of Law, Crime and Justice*, 37(3), pp. 83–103. doi: 10.1016/j.ijlcj.2009.04.002.

DeCew, J. (2015) 'Privacy', in Zalta, E. N. (ed.) *The Stanford Encyclopedia of Philosophy*. Spring 2015. Metaphysics Research Lab, Stanford University. Available at: https://plato.stanford.edu/archives/spr2015/entries/privacy/ (Accessed: 8 December 2016).

Dennet, D., C. (1996) in Stork, D. (ed.) *HAL's Legacy: '2001's' Computer as Dream and Reality*. New ed. edition. Cambridge, Mass: MIT Press.

Directorate General, Human Rights and Rule of Law and Council of Europe (2013) *Guide to Good Practice in Respect of Domestic Remedies*. Available at: https://www.coe.int/t/dghl/standardsetting/cddh/CDDH-DOCUMENTS/GuideBonnesPratiques-FINAL-EN.pdf.

Doan, S., Vo, B.-K. H. and Collier, N. (2011) 'An Analysis of Twitter Messages in the 2011 Tohoku Earthquake', in *Electronic Healthcare*. *International Conference on Electronic Healthcare*, Springer, Berlin, Heidelberg (Lecture Notes of the Institute for Computer

Sciences, Social Informatics and Telecommunications Engineering), pp. 58–66. doi: 10.1007/978-3-642-29262-0_8.

Donnelly, J. (2013) *Universal Human Rights in Theory and Practice*. 3rd Revised edition edition. Ithaca: Cornell University Press.

Dorasamy, M., Raman, M. and Kaliannan, M. (2013) 'Knowledge management systems in support of disasters management: A two decade review', *Technological Forecasting and Social Change*. (Planning and Foresight Methodologies in Emergency Preparedness and Management), 80(9), pp. 1834–1853. doi: 10.1016/j.techfore.2012.12.008.

Elliott, J. R. and Pais, J. (2006) 'Race, class, and Hurricane Katrina: Social differences in human responses to disaster', *Social Science Research*, 35(2), pp. 295–321. doi: 10.1016/j.ssresearch.2006.02.003.

Enarson, E. (2014) 'Human security and disasters: what a gender lens offers', in Hobson, C., Bacon, P., and Cameron, R. (eds) *Human Security and Natural Disasters*. 1st edn. New York, NY, USA: Routledge, pp. 37–56.

Ess, C. (2009) 'Floridi's philosophy of information and information ethics: Current perspectives, future directions', *The information society*, 25(3), pp. 159–168.

Ess, C. M. (2010) 'Trust and New Communication Technologies: Vicious Circles, Virtuous Circles, Possible Futures', *Knowledge, Technology and Policy*, 23(3–4), pp. 287–305.

European Commission (no date) *What is FP7? - FP7 in Brief - Research - EUROPA*. Available at: https://ec.europa.eu/research/fp7/understanding/fp7inbrief/what-is_en.html (Accessed: 24 July 2017).

European Union Agency for Fundamental Rights and Council of Europe (2010) *Handbook on European non-discrimination law*. 1st edn. Luxembourg: European Union Agency for Fundamental Rights European Court of Human Rights - Council of Europe.

Farhi, P. (2012) 'Twitter rumor of flooded stock-exchange among a surge of fake storm reports', *The Washington Post*, 30 October. Available at: https://www.washingtonpost.com/lifestyle/style/twitter-rumor-of-flooded-stock-exchange-among-a-surge-of-fake-storm-reports/2012/10/30/c03acc2c-22ce-11e2-8448-81b1ce7d6978_story.html?utm_term=.ce3ba2cc9d54 (Accessed: 8 June 2017).

Federal Emergency Management Agency (no date) 'Emergency Management Definition, Vision, Mission'. Available at: https://training.fema.gov/hiedu/docs/emprinciples/0907_176%20em%20principles12x18v2f%20johnson%20(w-o%20draft).pdf (Accessed: 2 August 2017).

Feinberg, J. (1985) 'Ethical Issues in the Use of Computers', in Johnson, D. G. and Snapper, J. W. (eds). Belmont, CA, USA: Wadsworth Publ. Co., pp. 102–120. Available at: http://dl.acm.org/citation.cfm?id=2569.2675.

Floridi, L. (2002) 'Information Ethics: An Environmental Approach to the Digital Divide', *Philosophy in the Contemporary World*, 9(1), pp. 39–45. doi: 10.5840/pcw2002915.

Floridi, L. (2005) 'The Ontological Interpretation of Informational Privacy', *Ethics and Information Technology*, 7, pp. 185–200.

Floridi, L. (2006) 'Four challenges for a theory of informational privacy', *Ethics and Information Technology*, 8(3), pp. 109–119. doi: 10.1007/s10676-006-9121-3.

Floridi, L. (2008) 'A Defence of Informational Structural Realism', *Synthese*, 161, pp. 219–253.

Floridi, L. (2011a) 'A Defence of Constructionism: Philosophy as Conceptual Engineering', *Metaphilosophy*, 42(3), pp. 282–304.

Floridi, L. (2011b) *The Philosophy of Information*. Oxford ; New York: OUP Oxford.

Floridi, L. (2013) *The Ethics of Information*. Oxford: OUP Oxford.

Floridi, L. (2014) *The Fourth Revolution: How the Infosphere is Reshaping Human Reality*. OUP Oxford.

Foddy, W. (1994) *Constructing Questions for Interviews and Questionnaires: Theory and Practice in Social Research: Written by William Foddy, 1994 Edition,*. Cambridge University Press.

Fothergill, A., Maestas, E. G. and Darlington, J. D. (1999) 'Race, ethnicity and disasters in the United States: a review of the literature', *Disasters*, 23(2), pp. 156–173.

Fox, R. (2001) 'Someone to Watch Over Us:: Back to the Panopticon?', *Criminal Justice*, 1(3), pp. 251–276. doi: 10.1177/1466802501001003001.

Fox-Decent, E. (2011) *Sovereignty's Promise: The State as Fiduciary*. Oxford ; New York: OUP Oxford.

Fox-Decent, E. and Criddle, E. (2012) 'Interest-Balancing vs. Fiduciary Duty: Two Models for National Security Law', *13 German Law Journal 542-559 (2012)*. Available at: http://scholarship.law.wm.edu/facpubs/1532.

Fox-Decent, E. and Criddle, E. J. (2010) 'The Fiduciary Constitution of Human Rights', *Legal Theory*, 15, pp. 301–336.

Fritz, C. (1961) 'Disasters', in Merton, R. K. and Nisbet, R. A. (eds) *Contemporary social problems: an introduction to the sociology of deviant behavior and social disorganization*. Harcourt, Brace & World, pp. 651–694.

Fuller, L. L. (1977) *The Morality of Law*. Revised edition edition. New Haven; London: Yale University Press.

Futamura, M., Hobson, C. and Turner, N. (2011) 'Natural Disasters and Human Security'. Available at: https://unu.edu/publications/articles/natural-disasters-and-human-security.html (Accessed: 2 August 2017).

Gomm, R., Hammersley, M. and Foster, P. (eds) (2006) *Case Study Method*. Norfolk, Britain: SAGE.

Gosepath, S. (2011) 'Equality', in Zalta, E. N. (ed.) *The Stanford Encyclopedia of Philosophy*. Spring 2011. Metaphysics Research Lab, Stanford University. Available at: https://plato.stanford.edu/archives/spr2011/entries/equality/ (Accessed: 6 April 2017).

Greco, G. M. *et al.* (2005) 'How to Do Philosophy Informationally', in *Professional Knowledge Management*. *Biennial Conference on Professional Knowledge Management/Wissensmanagement*, Springer, Berlin, Heidelberg (Lecture Notes in Computer Science), pp. 623–634. doi: 10.1007/11590019_70.

Greenwood, S., Perrin, A, and Duggan, M. (2016) 'Social Media Update 2016', *Pew Research Center: Internet, Science & Tech*, 11 November. Available at: http://www.pewinternet.org/2016/11/11/social-media-update-2016/ (Accessed: 6 April 2017).

Grodzinsky, F. S., Miller, K. W. and Wolf, M. J. (2011) 'Developing Artificial Agents Worthy of Trust: "Would You Buy a Used Car from This Artificial Agent?"', *Ethics and Inf. Technol.*, 13(1), pp. 17–27. doi: 10.1007/s10676-010-9255-1.

Gross, O. (2008) 'Extra-legality and the ethic of political responsibility', in Ramraj, V. V. (ed.) *Emergencies and the Limits of Legality*. 1 edition. Cambridge: Cambridge University Press, pp. 60–93.

Gross, O. and Ni Aolain, F. (2006) *Law in Times of Crisis: Emergency Powers in Theory and Practice*. 1 edition. Cambridge: Cambridge University Press.

Guillen, M. F. and Suarez, S. L. (2006) 'Explaining the Global Digital Divide: Economic, Political and Sociological Drivers of Cross-National Internet Use', *Social Forces*, 84(2), pp. 681–708. doi: 10.1353/sof.2006.0015.

Gupta, A. *et al.* (2013) 'Faking Sandy: Characterizing and Identifying Fake Images on Twitter During Hurricane Sandy', in *Proceedings of the 22Nd International Conference on World Wide Web*. New York, NY, USA: ACM (WWW '13 Companion), pp. 729–736. doi: 10.1145/2487788.2488033.

Haase, T., Pratschke, J. and Gleeson, J. (2014) *AIRO-2011-All-Island-HP-Deprivation-Index.jpg (3533×2505)*. Available at: http://trutzhaase.eu/wp/wp-content/uploads/AIRO-2011-All-Island-HP-Deprivation-Index.jpg (Accessed: 16 April 2017).

Hafner-Burton, E. M., Helfer, L. R. and Fariss, C. J. (2011) 'Emergency and Escape: Explaining Derogations from Human Rights Treaties', *International Organization*, 65(4), pp. 673–707. doi: 10.1017/S002081831100021X.

Hardwig, J. (1985) ‘Epistemic Dependence’, *Journal of Philosophy*, 82(7), pp. 335–349.

Harris, D. *et al.* (2009) *Harris, O’Boyle & Warbrick: Law of the European Convention on Human Rights*. 2 edition. Oxford ; New York: OUP Oxford.

Hayes, P. (2017) ‘Information without borders: towards a framework for extra-territorial respect for the right to privacy’, *International Journal of Human Rights and Constitutional Studies*, 5(1), pp. 60–81. doi: 10.1504/IJHRCS.2017.10003666.

Hellström, T. (2013) ‘On the moral responsibility of military robots’, *Ethics and Information Technology*, 15(2), pp. 99–107. doi: 10.1007/s10676-012-9301-2.

Himma, K. E. (2004) ‘There’s something about Mary: The moral value of things qua information objects’, *Ethics and Information Technology*, 6(3), pp. 145–159. doi: 10.1007/s10676-004-3804-4.

Himma, K. E. (2009) ‘Artificial agency, consciousness, and the criteria for moral agency: what properties must an artificial agent have to be a moral agent?’, *Ethics and Information Technology*, 11(1), pp. 19–29. doi: 10.1007/s10676-008-9167-5.

Hobson, C., Bacon, P. and Cameron, R. (eds) (2014) *Human Security and Natural Disasters*. London ; New York: Routledge.

van den Hoven, J. (1998) ‘Moral responsibility, public office and information technology’, in Snellen, I. T. M. and Donk, W. V. D. (eds) *Public Administration in an Information Age, A Handbook*. Amsterdam Berlin: Ios Pr Inc, pp. 97–112.

Hughes, A., L. *et al.* (2014) ‘Online public communications by police & fire services during the 2012 Hurricane Sandy’, in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Toronto, Ontario, Canada: ACM, pp. 1505–1514. Available at: http://s3.amazonaws.com/academia.edu.documents/34676306/HughesStDenisPalenAndersonPoliceFireSandy.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1501722912&Signature=mEhddVLbqdKTBZGnNTLl4S64Zsk%3D&response-content-disposition=inline%3B%20filename%3DHughes_St_Denis_Palen_Anderson_Police_Fi.pdf (Accessed: 2 August 2017).

Human Rights Watch (2014) ‘“They Know Everything We Do” Telecom and Internet Surveillance in Ethiopia’. Available at: https://www.hrw.org/report/2014/03/25/they-know-everything-we-do/telecom-and-internet-surveillance-ethiopia (Accessed: 11 December 2016).

Hunt, E. (2016) ‘What is fake news? How to spot it and what you can do to stop it’, *The Guardian*, 17 December. Available at: https://www.theguardian.com/media/2016/dec/18/what-is-fake-news-pizzagate.

Imran, M. *et al.* (2014) ‘AIDR: Artificial Intelligence for Disaster Response’, in *IW3C2*. *International World Wide Web Conference*, Seoul, Korea.

International Federation of Red Cross and Red Crescent Societies (no date) 'What is a disaster?' Available at: https://www.ifrc.org/en/what-we-do/disaster-management/about-disasters/what-is-a-disaster/ (Accessed: 2 August 2017).

International Federation of Red Cross and Red Crescent Societies (2013) *World Disasters Report 2013*. Available at: http://www.ifrc.org/PageFiles/134658/WDR%202013%20complete.pdf (Accessed: 11 December 2016).

Jackson, D., Aldrovandi, C. and Hayes, P. (2015) 'Ethical Framework for a Disaster Management Decision Support System Which Harvests Social Media Data on a Large Scale', in Saoud, N. B. B., Adam, C., and Hanachi, C. (eds) *Information Systems for Crisis Response and Management in Mediterranean Countries*. Springer International Publishing, pp. 167–180. Available at: http://link.springer.com/chapter/10.1007/978-3-319-24399-3_15 (Accessed: 23 September 2016).

Jansson, A. (2012) 'Perceptions of surveillance: Reflexivity and trust in a mediatized world (the case of Sweden)', *European Journal of Communication*, 27(4), pp. 410–427. doi: 10.1177/0267323112463306.

Jennex, M., E. (2012) 'Social media - Truly viable for crisis response?', in Franco, Z. and Rothkrantz, J., R, L. (eds) *ISCRAM 2012*. *9th International Conference on Information Systems for Crisis Response and Management*, Vancouver, BC: Simon Fraser University, pp. 53–67. Available at: https://www.academia.edu/31038986/Social_Media_Viable_for_Crisis_Response_Experience_from_the_Great_San_Diego_Southwest_Blackout (Accessed: 2 August 2017).

Johnson, D. G. and Mulvey, J. M. (1995) 'Accountability and Computer Decision Systems', *Commun. ACM*, 38(12), pp. 58–64. doi: 10.1145/219663.219682.

Johnson, D. G. and Powers, T. M. (2005) 'Computer Systems and Responsibility: A Normative Look at Technological Complexity', *Ethics and Information Technology*, 7(2), pp. 99–107. doi: 10.1007/s10676-005-4585-0.

Johnstone, J. (2007) 'Technology as empowerment: a capability approach to computer ethics', *Ethics and Information Technology*, 9(1), pp. 73–87. doi: 10.1007/s10676-006-9127-x.

Joyner, C. C. (1997) 'Redressing Impunity for Human Rights Violations: The Universtal Declaration and the Search for Accountability', *Denver Journal of International Law and Policy*, 26, p. 591.

Kelly, T. *et al.* (2014) *Cork City Profile 2014: Section I: A Statistical and Geographical Profile of Cork City Local Authority Area Focused on Health and Social Inclusion*. Cork Corporation Bardas Clorcat.

Kim, N. (2012) 'How much more exposed are the poor to natural disasters? Global and regional measurement', *Disasters*, 36(2), pp. 195–211. doi: 10.1111/j.1467-7717.2011.01258.x.

King, G. and Murray, C. J. L. (2001) 'Rethinking Human Security', *Political Science Quarterly*, 116(4), pp. 585–610. doi: 10.2307/798222.

King, H. (2009) 'The Extraterritorial Human Rights Obligations of States', *Human Rights Law Review*, 9(4), pp. 521–556.

Korff, D. (2006) *The right to life: A guide to implementation of Article 2 of the European Convention on Human Rights*. 1st edn. Belgium: Council of Europe.

Lasén, A. and Gómez-Cruz, E. (2009) 'Digital Photography and Picture Sharing: Redefining the Public/Private Divide', *Knowledge, Technology & Policy*, 22(3), pp. 205–215. doi: 10.1007/s12130-009-9086-8.

Latonero, M. and Shklovski, I. (2011) 'Emergency management, Twitter, & Social Media Evangelism', *International Journal of Information Systems for Crisis Response and Management*, 3(4), pp. 67–86.

Lazar, N., C. (2008) 'A topography of emergency power', in Ramraj, V. V. (ed.) *Emergencies and the Limits of Legality*. 1 edition. Cambridge: Cambridge University Press, pp. 156–171.

Lomas, N. (2017) 'Trump order strips privacy rights from non-U.S. citizens, could nix EU-US data flows', *TechCrunch*. Available at: http://social.techcrunch.com/2017/01/26/trump-order-strips-privacy-rights-from-non-u-s-citizens-could-nix-eu-us-data-flows/ (Accessed: 5 February 2017).

Magnani, L. (2005) 'Moral Mediators', in Magnani, L. and Dossena, R. (eds) *Computing, Philosophy and Cognition*. London: College Publications, pp. 1–16.

Magnani, L. and Bardone, E. (2008) 'Distributed Morality: Externalizing Ethical Knowledge in Technological Artifacts', *Foundations of Science*, 13, p. pp.99-108.

Marder, B. *et al.* (2016) 'The extended "chilling" effect of Facebook: The cold reality of ubiquitous social networking', *Computers in Human Behavior*, 60, pp. 582–592. doi: 10.1016/j.chb.2016.02.097.

Margulies, P. (2014) 'The NSA in Global Perspective: Surveillance, Human Rights, and International Counterterrorism', *Fordham Law Review*, 82(5), pp. 2137–2167.

Masozera, M., Bailey, M. and Kerchner, C. (2007) 'Distribution of impacts of natural disasters across income groups: A case study of New Orleans', *Ecological Economics*, 63(2–3), pp. 299–306. doi: 10.1016/j.ecolecon.2006.06.013.

McLeod, C. (2015) ‘Trust’, in Zalta, E. N. (ed.) *The Stanford Encyclopedia of Philosophy*. Fall 2015. Metaphysics Research Lab, Stanford University. Available at: https://plato.stanford.edu/archives/fall2015/entriesrust/.

Mendoza, M., Poblete, B. and Castillo, C. (2010) ‘Twitter Under Crisis: Can We Trust What We RT?’, in *Proceedings of the First Workshop on Social Media Analytics*. New York, NY, USA: ACM (SOMA ’10), pp. 71–79. doi: 10.1145/1964858.1964869.

Menon, R., R. (2010) ‘Natural Hazards and Unnatural Disasters: A Survey of the Gendered Terrain of Risk, Vulnerability, and Disaster Relief’, in Fuentes-Nieva, R. and Seck, P., A. (eds) *Risks, Shocks, and Human Development On the Brink*. 1st edn. Basingstoke, England ; New York, pp. 310–341.

Milanovic, M. (2015) ‘Human Rights Treaties and Foreign Surveillance: Privacy in the Digital Age’, *Harvard International Law Journal*, 56(1), pp. 81–145.

Mirriam-Webster (no date) *Definition of BETRAY*. Available at: https://www.merriam-webster.com/dictionary/betray (Accessed: 21 September 2017).

Mitchell, J. C. (2006) ‘Case and Situation Analysis’, in Gomm, R., Hammersley, M., and Foster, P. (eds) *Case Study Method*. Norfolk, Britain: SAGE, pp. 165–186.

mixi (no date) *About Social Network - mixi, Inc.*, *About Social Network*. Available at: http://mixi.co.jp/en/about/ (Accessed: 2 August 2017).

Moor, J. H. (1990) ‘The Ethics of Privacy Protection’, *Library Trends*, 39(1–2), pp. 69–82.

Moor, J. H. (1997) ‘Towards a Theory of Privacy in the Information Age’, *SIGCAS Computers and Society*, 27(3), pp. 27–32. doi: 10.1145/270858.270866.

Moor, J. H. (2006) ‘The Nature, Importance, and Difficulty of Machine Ethics’, *IEEE Intelligent Systems*, p. pp.18-21.

Mordini, E. *et al.* (2009) ‘Senior citizens and the ethics of e-inclusion’, *Ethics and Information Technology*, 11(3), pp. 203–220. doi: 10.1007/s10676-009-9189-7.

Morrow, N. *et al.* (2011) ‘Independent Evaluation of the Ushahidi Haiti Project’. Available at: http://www.alnap.org/pool/files/1282.pdf (Accessed: 2 August 2017).

Moss, J. (2002) ‘“Power and the digital divide”’, *Ethics and Information Technology*, 4(2), pp. 159–165. doi: 10.1023/A:1019983909305.

Mowbray, A. (2002) ‘Duties of Investigation Under the European Convention on Human Rights’, *International & Comparative Law Quarterly*, 51(2), pp. 437–448. doi: 10.1093/iclq/51.2.437.

de Mul, J. (2010) ‘Moral Machines: ICTs as Mediators of Human Agency’, *Techné*, 14(3), pp. 226–236.

Munich Re (2015) 'NatCatSERVICE Loss events worldwide 1980 – 2014'. Available at: https://www.munichre.com/site/touch-naturalhazards/get/documents_E2080665585/mr/assetpool.shared/Documents/5_Touch/_NatCatService/Focus_analyses/1980-2014-Loss-events-worldwide.pdf (Accessed: 2 August 2017).

Nel, P. and Righarts, M. (2008) 'Natural Disasters and the Risk of Violent Civil Conflict', *International Studies Quarterly*, 52(1), pp. 159–185. doi: 10.1111/j.1468-2478.2007.00495.x.

Neumayer, E. and Plümper, T. (2007) 'The gendered nature of natural disasters: the impact of catastrophic events on the gender gap in life expectancy, 1981–2002', *Annals of the Association of American Geographers*, 97(3), pp. 551–566.

Nissenbaum, H. F. (1996) 'Accountability in a Computerized Society', *Science and Engineering Ethics*, 2, pp. 25–42.

Nissenbaum, H. F. (2001) 'Securing Trust Online: Wisdom or Oxymoron?', *Boston University Law Review*, 81, pp. 635–664.

Nissenbaum, H. F. (2009) *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford, Calif: Stanford University Press.

Noorman, M. (2016) 'Computing and Moral Responsibility', in Zalta, E. N. (ed.) *The Stanford Encyclopedia of Philosophy*. Winter 2016. Metaphysics Research Lab, Stanford University. Available at: https://plato.stanford.edu/archives/win2016/entries/computing-responsibility/.

Nussbaum, M. (2003) 'Capabilities as Fundamental Entitlements: Sen and Social Justice', *Feminist Economics*, 9(2–3), pp. 33–59. doi: 10.1080/1354570022000077926.

Nussbaum, M. (2008) *PCBE: Human Dignity and Bioethics:Essays Commissioned by the President's Council on Bioethics (Chapter 14: Human Dignity and Political entitlements)*. Available at: https://bioethicsarchive.georgetown.edu/pcbe/reports/human_dignity/chapter14.html (Accessed: 6 April 2017).

O'Brien, M. (2008) 'Law, privacy and information technology: a sleepwalk through the surveillance society?', *Information & Communications Technology Law*, 17(1), pp. 25–35. doi: 10.1080/13600830801887214.

OECD (2001) *Understanding the Digital Divide*. Paris, France. Available at: https://www.oecd.org/sti/1888451.pdf (Accessed: 6 April 2017).

Oh, O., Agrawal, M. and Rao, H. R. (2013) 'Community Intelligence and Social Media Services: A Rumor Theoretic Analysis of Tweets During Social Crises', *MIS Q.*, 37(2), pp. 407–426.

Oxford Dictionaries (no date) *entrust - definition of entrust in English | Oxford Dictionaries*, *Oxford Dictionaries | English*. Available at: https://en.oxforddictionaries.com/definition/entrust (Accessed: 21 June 2017).

Ozturk, P., Li, H. and Sakamoto, Y. (2015) 'Combating Rumor Spread on Social Media: The Effectiveness of Refutation and Warning', in *2015 48th Hawaii International Conference on System Sciences*. *2015 48th Hawaii International Conference on System Sciences*, pp. 2406–2414. doi: 10.1109/HICSS.2015.288.

Paquette, S. and Yates, D. (2011) 'Emergency knowledge management and social media technologies: A case study of the 2010 Haitian earthquake', *International Journal of Information Management*, (31), pp. 6–13.

Parfit, D. (1997) 'Equality and Priority', *Ratio*, 10(3), pp. 202–221. doi: 10.1111/1467-9329.00041.

Parilla-Ferrer, B., E., Fernandez Jr., P., L. and Balena, J., T. (2014) 'Automatic Classification of Disaster-Related Tweets', in *ICIET'2014*. *International conference on Innovative Engineering Technologies*, Bangkok, Thailand. Available at: https://www.researchgate.net/publication/282154924_Automatic_Classification_of_Disaster-Related_Tweets (Accessed: 2 August 2017).

Peary, B., D., Shaw, R. and Takeuchi, Y. (2012) 'Utilization of Social Media in the East Japan Earthquake and Tsunami and its Effectiveness', *Journal of Natural Disaster Science*, 34(1), pp. 3–18.

Penney, J. (2016) 'Chilling Effects: Online Surveillance and Wikipedia Use', *Berkeley Technology Law Journal*, 31(1), p. 117. doi: http://dx.doi.org/10.15779/Z38SS13.

Perrin, A. (2015) 'Social Media Usage: 2005-2015', *Pew Research Center: Internet, Science & Tech*, 8 October. Available at: http://www.pewinternet.org/2015/10/08/social-networking-usage-2005-2015/ (Accessed: 6 April 2017).

Phillips, B. D., Neal, D. M. and Webb, G. (2011) *Introduction to Emergency Management*. 1 edition. Boca Raton, FL: CRC Press.

Pitkin, H. (1966) 'Obligation and Consent--II', *The American Political Science Review*, 60(1), pp. 39–52. doi: 10.2307/1953805.

Piven, B. (2016) *Sunil's saga: Family double-victimized by depression, Internet witch hunt*. Available at: http://america.aljazeera.com/articles/2016/2/12/sunil-family-depression-victimization.html (Accessed: 9 June 2017).

Popovski, V. (2014) 'State negligence before and after natural disasters as human rights violations', in Hobson, C., Bacon, P., and Cameron, R. (eds) *Human Security and Natural Disasters*. London ; New York: Routledge, pp. 94–110.

Popovski, V. (2016) 'State negligence before and after natural disasters as human rights violations', in Hobson, C., Bacon, P., and Cameron, R. (eds) *Human Security and Natural Disasters*. London ; New York: Routledge, pp. 94–110.

Poushter, J. (2016) 'Smartphone Ownership and Internet Usage Continues to Climb in Emerging Economies', *Pew Research Center's Global Attitudes Project*, 22 February. Available at: http://www.pewglobal.org/2016/02/22/smartphone-ownership-and-internet-usage-continues-to-climb-in-emerging-economies/ (Accessed: 6 April 2017).

Protalinski, E. (2014) 'Facebook passes 1.23 billion monthly active users, 945 million mobile users, and 757 million daily users'. Available at: https://thenextweb.com/facebook/2014/01/29/facebook-passes-1-23-billion-monthly-active-users-945-million-mobile-users-757-million-daily-users/#.tnw_jC6YcqZZ (Accessed: 2 August 2017).

Purenne, A. and Palierse, G. (2017) 'Towards Cities of Informers? Community-Based Surveillance in France and Canada', *Surveillance & Society*, 15(1), pp. 79–93.

Quarantelli, E. L. (2008) 'Conventional Beliefs and Counterintuitive Realities', *Social Research: An International Quarterly*, 75(3), pp. 873–904.

Ranghieri, F. and Ishiwatari, M. (eds) (2014) *Learning from Megadisasters: Lessons from the Great East Japan Earthquake*. Washington, DC: World Bank Publications.

Rawls, J. (1958) 'Justice as Fairness', *The Philosophical Review*, 67(2), pp. 164–194. doi: 10.2307/2182612.

Rooksby, E. (2009) 'How to be a responsible slave: managing the use of expert information systems', *Ethics and Information Technology*, 11(1), pp. 81–90. doi: 10.1007/s10676-009-9183-0.

Rousseau, J.-J. (1974) *The Essential Rousseau*. Translated by L. Bair. New York, N.Y.: Plume/Meridian.

Sahana Foundation (2015) 'Map'. Available at: http://demo.eden.sahanafoundation.org/eden/gis/index (Accessed: 2 August 2017).

Scheinin, M. (2009) *Report of the Special Rapporteur on the promotion and protection of human rights and fundamental freedoms while countering terrorism*. A/HRC/13/37. Available at: http://www2.ohchr.org/english/bodies/hrcouncil/docs/13session/A-HRC-13-37.pdf.

Schmitt, C. and Strong, T. B. (2005) *Political Theology: Four Chapters on the Concept of Sovereignty*. University of Chicago Press Ed edition. Edited by G. Schwab. Chicago: University of Chicago Press.

Sen, A. (2001) *Development as Freedom*. New Ed edition. Oxford ; New York: OUP Oxford.

Seo, E., Mohapatra, P. and Abdelzaher, T. (2012) 'Identifying rumors and their sources in social networks', in, p. 83891I–83891I–13. doi: 10.1117/12.919823.

Sheeran, C. (2015) *How are Irish Travellers experiencing and addressing the 'digital divide' in Ireland*. Dublin City University. Available at: https://www.academia.edu/30918496/How_are_Irish_Travellers_experiencing_and_addressing_the_digital_divide_in_Ireland (Accessed: 20 April 2017).

Shubber, K. (2013) *A simple guide to GCHQ's internet surveillance programme Tempora (Wired UK)*, *Wired UK*. Available at: http://www.wired.co.uk/news/archive/2013-06/24/gchq-tempora-101 (Accessed: 18 May 2016).

Silverman, D. (2013) *Doing Qualitative Research: A Practical Handbook*. 4th edn. Great Britain: SAGE Publications Ltd.

Simon, J. (2015) 'Distributed Epistemic Responsibility in a Hyperconnected Era', in Floridi, L. (ed.) *The Onlife Manifesto: Being Human in a Hyperconnected Era*. 1st edn. Springer, p. pp.145-159. Available at: http://www.springer.com/gp/book/9783319040929 (Accessed: 29 November 2016).

Smiley, M. (2017) 'Collective Responsibility', in Zalta, E. N. (ed.) *The Stanford Encyclopedia of Philosophy*. Summer 2017. Metaphysics Research Lab, Stanford University. Available at: https://plato.stanford.edu/archives/sum2017/entries/collective-responsibility/.

Sociological Association of Ireland (no date) 'Ethical Guidelines of the Sociological Association of Ireland'. Available at: http://www.sociology.ie/uploads/4/2/5/2/42525367/sai_ethical_guidelines.pdf (Accessed: 20 September 2017).

Sommario, E. (2012) 'Chapter 14: Derogations from Human Rights Treaties in Situations of Natural or Man-Made Disasters', in de Grutty, A., Gestri, M., and Venturini, G. (eds) *International Disaster Response Law*. 1st edn. The Hague, Netherlands: Asser Press, pp. 323–351.

Stahl, B. C. (2006a) 'Accountability and Reflective Responsibility in Information Systems', in *The Information Society: Emerging Landscapes*. Springer, Boston, MA (IFIP International Federation for Information Processing), pp. 51–68. Available at: https://pdfs.semanticscholar.org/cf99/8c5840faf1a9f9e39c3a6372ee214a4ddc86.pdf (Accessed: 14 July 2017).

Stahl, B. C. (2006b) 'Responsible computers? A case for ascribing quasi-responsibility to computers independent of personhood or agency', *Ethics and Information Technology*, 8(4), pp. 205–213. doi: 10.1007/s10676-006-9112-4.

Starbird, K. *et al.* (2010) 'Chatter on the red: what hazards threat reveals about the social life of microblogged information', in *Proceedings of the 2010 ACM conference on Computer supported cooperative work*. Savannah, Georgia, USA: ACM, pp. 241–250.

Starbird, K. *et al.* (2014) 'Rumors, False Flags, and Digital Vigilantes: Misinformation on Twitter after the 2013 Boston Marathon Bombing', in. *iConference 2014*, Berlin, Germany. doi: 10.9776/14308.

Starbird, K. and Palen, L. (2010) 'Pass It On?: Retweeting in Mass Emergency', in *Proceedings of the 7th International ISCRAM Conference*. *ISCRAM 2010*, Seattle, USA. Available at: http://faculty.washington.edu/kstarbi/iscram_Retweet_FinalPaper.pdf (Accessed: 2 August 2017).

Statista (2017a) *Leading countries based on number of Facebook users as of July 2017 (in millions)*, *Statista*. Available at: https://www.statista.com/statistics/268136/top-15-countries-based-on-number-of-facebook-users/ (Accessed: 2 August 2017).

Statista (2017b) *Number of monthly active Facebook users worldwide as of 2nd quarter 2017 (in millions)*, *Statista*. Available at: https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/ (Accessed: 3 August 2017).

Statista (2017c) *Number of monthly active Twitter users worldwide from 1st quarter 2010 to 2nd quarter 2017 (in millions)*, *Statista*. Available at: https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/ (Accessed: 2 August 2017).

Statista (2017d) *Number of social media users worldwide from 2010 to 2021 (in billions)*, *Statista*. Available at: https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/ (Accessed: 2 August 2017).

Steiner, P. and The New Yorker Magazine (1993) *On the Internet, nobody knows you're a dog*. Available at: https://img.washingtonpost.com/rf/image_1484w/WashingtonPost/Content/Blogs/comic-riffs/StandingArt/STEINERinternetdogs.jpg?uuid=Cn7v6vmREeKOhMVnMaIC-w (Accessed: 14 June 2017).

Stroud, S. (2014) 'Weakness of Will', in Zalta, E. N. (ed.) *The Stanford Encyclopedia of Philosophy*. Spring 2014. Metaphysics Research Lab, Stanford University. Available at: https://plato.stanford.edu/archives/spr2014/entries/weakness-will/.

Suebsaeng, A. (2013) 'My Innocent Brother Was Made Into a Bombing Suspect: Sunil Tripathi's Sister Speaks', *Mother Jones*. Available at: http://www.motherjones.com/politics/2013/04/sunil-tripathi-sister-sangeeta-media-labelling-her-brother-bombing-suspect/ (Accessed: 9 June 2017).

Sutton, J., Palen, L. and Shklovsk, I. (2008) 'Backchannels on the Front Lines: Emergent Uses of Social Media in the 2007 Southern California Wildfires', in *e 5th International ISCRAM Conference*. Washington, DC, USA. Available at: http://jeannettesutton.com/uploads/BackchannelsISCRAM08.pdf (Accessed: 6 December 2016).

Taddeo, M. (2009) 'Defining Trust and E-Trust: From Old Theories to New Problems', *International Journal of Technology and Human Interaction (IJTHI)*, 5(2), pp. 23–35. doi: 10.4018/jthi.2009040102.

Taddeo, M. (2010a) 'An Information-based Solution for the Puzzle of Testimony and Trust', *Social Epistemology*, 24(4), pp. 285–299. doi: 10.1080/02691728.2010.521863.

Taddeo, M. (2010b) 'Modelling Trust in Artificial Agents, A First Step Toward the Analysis of e-Trust', *Minds and Machines*, 20(2), pp. 243–257. doi: 10.1007/s11023-010-9201-3.

Takayasu, M. *et al.* (2015) 'Rumor Diffusion and Convergence during the 3.11 Earthquake: A Twitter Case Study', *PLOS ONE*, 10(4), p. e0121443. doi: 10.1371/journal.pone.0121443.

Tavani, H. T. (2007) 'Philosophical Theories of Privacy: Implications for an Adequate Online Privacy Policy', *Metaphilosophy*, 38(1), pp. 1–22.

Tavani, H. T. (2008) 'Floridi's ontological theory of informational privacy: Some implications and challenges', *Ethics and Information Technology*, 10(2–3), pp. 155–166. doi: 10.1007/s10676-008-9154-x.

The Hamilton Project at the Brookings Institution (2011) *Cost of Computing Power Equal to an iPad2*. Available at: http://www.hamiltonproject.org/charts/cost_of_computing_power_equal_to_an_ipad2/ (Accessed: 2 August 2017).

Thompson, S. *et al.* (2006) 'Improving Disaster Response Efforts With Decision Support Systems', *International Journal of Emergency Management*, 3(4), pp. 250–263. doi: 10.1504/IJEM.2006.011295.

Thoms, O. N. T. and Ron, J. (2007) 'Do Human Rights Violations Cause Internal Conflict?', *Human Rights Quarterly*, 29(3), pp. 674–705. doi: 10.1353/hrq.2007.0034.

Tognazzini, N. and Coates, D. J. (2016) 'Blame', in Zalta, E. N. (ed.) *The Stanford Encyclopedia of Philosophy*. Spring 2016. Metaphysics Research Lab, Stanford University. Available at: https://plato.stanford.edu/archives/spr2016/entries/blame/.

Trainor, J., Barsky, L. and Torres, M. (2006) *Disaster Realities in the Aftermath of Hurricane Katrina: Revisiting the Looting Myth*. Miscellaneous Report 184. Available at: http://udspace.udel.edu/handle/19716/2367 (Accessed: 7 June 2017).

Trinity College Dublin (2002) 'Good Research Practice'. Available at: https://www.tcd.ie/Graduate_Studies/assets/pdfs/TCD%20Good%20Research%20Practice.pdf (Accessed: 20 September 2017).

Tuckness, A. (2016) 'Locke's Political Philosophy', in Zalta, E. N. (ed.) *The Stanford Encyclopedia of Philosophy*. Spring 2016. Metaphysics Research Lab, Stanford University. Available at: https://plato.stanford.edu/archives/spr2016/entries/locke-political/.

Turilli, M. (2007) 'Ethical protocols design', *Ethics and Information Technology*, 9, p. pp.49-62.

Turilli, M. and Floridi, L. (2009) 'The ethics of information transparency', *Ethics of Information Technology*, 11, pp. 105–112.

Turilli, M., Vaccaro, A. and Taddeo, M. (2010) 'The Case of Online Trust', *Knowledge, Technology & Policy*, 23(3–4), pp. 333–345. doi: 10.1007/s12130-010-9117-5.

Umihara, J. and Nishikitani (2013) 'Emergent Use of Twitter in the 2011 Tohoku Earthquake', *Prehospital and Disaster Medicine*, 28(5), pp. 434–440.

UN Commission on Human Rights (1984) *The Siracusa Principles on the Limitation and Derogation Provisions in the International Covenant on Civil and Political Rights*, *E/CN.4/1985/4*.

United Nations Development Programme (1994) 'Human Development Report 1994'. Available at: http://hdr.undp.org/en/content/human-development-report-1994 (Accessed: 2 August 2017).

United Nations Office of Disaster Risk Reduction (2012) *Number of Climate-related Disasters Around the World (1980-2011)*. Available at: http://www.preventionweb.net/files/20120613_ClimateDisaster1980-2011.pdf (Accessed: 2 August 2017).

United Nations Office of Disaster Risk Reduction (2013) *Disaster Impacts/2000-2012*. Available at: http://www.preventionweb.net/files/31737_20130312disaster20002012copy.pdf (Accessed: 2 August 2017).

United Nations Office of Disaster Risk Reduction (2017) *Terminology - UNISDR*. Available at: http://www.unisdr.org/we/inform/terminology#letter-d (Accessed: 2 August 2017).

Van Dyke, J. (2005) 'Promoting Accountability for Human Rights Abuses', *Chapman Law Review*, 8(1), pp. 148–171.

Van Schaak, B. (2014) 'The United States' Position on the Extraterritorial Application of Human Rights Obligations: Now is the Time for Change', *International Law Studies*, 90(20), pp. 20–65.

Vance, A., Lowry, P. B. and Eggett, D. (2013) 'Using Accountability to Reduce Access Policy Violations in Information Systems', *Journal of Management Information Systems*, 29(4), pp. 263–290. doi: 10.2753/MIS0742-1222290410.

Virtual Social Media Working Group and DHS First Responders Group (2013) 'Lessons Learned: Social Media and Hurricane Sandy'. Available at: https://www.dhs.gov/sites/default/files/publications/Lessons%20Learned%20Social%20Media%20and%20Hurricane%20Sandy.pdf (Accessed: 7 December 2013).

Vonnegut, K. (1994) *Welcome To The Monkey House and Palm Sunday: An Autobiographical Collage: 'Welcome to the Monkey House', 'Palm Sunday'*. London: Vintage Classics.

Warf, B. and Vincent, P. (2007) 'Multiple geographies of the Arab Internet', *Area*, 39(1), pp. 83–96. doi: 10.1111/j.1475-4762.2007.00717.x.

Warner, R. and Sloan, R. H. (2016) '"I'll See": How Surveillance Undermines Privacy By Eroding Trust', *Santa Clara High Technology Law Journal*, 32(2), p. 221.

Warren, S. D. and Brandeis, L. D. (1890) 'The Right to Privacy', *Harvard Law Review*, 4(5), pp. 193–220.

Weist, R., E., Mocellin, S. P. and Motsisi, D. T. (1994) *The Needs of Women in Disasters and Emergencies*. Winnipeg, Manitoba: Disaster Research Institute. Available at: https://www.gdnonline.org/resources/women-in-disaster-emergency.pdf (Accessed: 6 April 2017).

whitehouse.gov (2017) *Executive Order: Enhancing Public Safety in the Interior of the United States*, *whitehouse.gov*. Available at: https://www.whitehouse.gov/the-press-office/2017/01/25/presidential-executive-order-enhancing-public-safety-interior-united (Accessed: 5 February 2017).

Wiegel, V. (2010) 'The Ethics of IT-Artefacts', in Floridi, L. (ed.) *The Cambridge Handbook of Information and Computer Ethics,*. 1st edn. Cambridge, UK: Cambridge University Press, pp. 201–218.

Williams, M. L. *et al.* (2013) 'Policing cyber-neighbourhoods: tension monitoring and social media networks', *Policing and Society*, 23(4), pp. 461–481. doi: 10.1080/10439463.2013.780225.

World Bank (2017) *Internet users (per 100 people) | Data*. Available at: http://data.worldbank.org/indicator/IT.NET.USER.P2 (Accessed: 6 April 2017).

World Economic Forum (2013) *Digital Wildfires in a Hyperconnected World*, *Global Risks 2013*. Available at: http://wef.ch/GJCg5E (Accessed: 8 June 2017).

worldometers (2017) *World Population Clock: 7.5 Billion People (2017) - Worldometers*. Available at: http://www.worldometers.info/world-population/#table-forecast (Accessed: 3 August 2017).

Yang, F. *et al.* (2012) 'Automatic Detection of Rumor on Sina Weibo', in *MDS'12*. Available at: http://wan.poly.edu/KDD2012/forms/workshop/MDS12/doc/mds2012_submission_17.pdf (Accessed: 7 June 2017).

Zack, N. (2010) *Ethics for Disaster*. Rowman & Littlefield Publishers.

Zack, N. (2016) 'The ethics of disaster and Hurricane Katrina: human security, Homeland Security, and women's groups', in Hobson, C., Bacon, P., and Cameron, R. (eds) *Human Security and Natural Disasters*. 1st edn. New York, NY, USA: Routledge, pp. 57–73.

Case Law

Andreou v. Turkey [2010] (ECtHR). Available at: http://hudoc.echr.coe.int/ENG?i=001-95295 (Accessed:: 12 September 2017).

Appleby and Others v The United Kingdom [2003] (ECtHR). Available at: http://hudoc.echr.coe.int/eng?i=001-61080 (Accessed: 12 September 2017).

Axel Springer v Germany [2012] (ECtHR). Available at: http://hudoc.echr.coe.int/eng?i=001-109034 (Accessed: 12 September 2017).

Baczkowski and Others v. Poland [2007] (ECtHR). Available at: http://hudoc.echr.coe.int/eng?i=001-80464 (Accessed: 12 September 2017).

Bankovic v. Belgium and 16 Other Contracting Parties [1999] (ECtHR). Available at: http://hudoc.echr.coe.int/eng?i=001-22099 (Accessed: 12 September 2017).

Belpietro v. Italy [2013] (ECtHR). Available at: http://hudoc.echr.coe.int/eng?i=001-126450 (Accessed: 12 September 2017).

Boyle and Rice v. The United Kingdom, [1988] (ECtHR). Available at: http://hudoc.echr.coe.int/eng?i=001-57446 (Accessed: 12 September 2017).

Bubbins v. the United Kingdom [2006] (ECtHR). Available at: http://hudoc.echr.coe.int/eng?i=001-68548 (Accessed: 12 September 2017).

Budayeva and Others v. Russia [2008] (ECtHR). Available at: https://hudoc.echr.coe.int/eng#{"itemid":["001-85436"]} (Accessed: 12 September 2017).

D.H. and Others v. The Czech Republic [2007] (ECtHR). Available at: http://hudoc.echr.coe.int/eng?i=001-83256 (Accessed: 12 September 2017).

Eon v. France [2013] (ECtHR). Available at: http://hudoc.echr.coe.int/eng?i=001-117742 (Accessed: 12 September 2017).

Ezelin v. France [1991] (ECtHR). Available at: http://hudoc.echr.coe.int/eng?i=001-57675 (Accessed: 12 September 2017).

Gunduz v. Turkey [2004] (ECtHR). Available at: http://hudoc.echr.coe.int/eng?i=001-61522 (Accessed: 12 September 2017).

Jasinskis v. Latvia [2011] (ECtHR). Available at: http://hudoc.echr.coe.int/eng?i=001-102393 (Accessed: 12 September 2017).

Khan v. The United Kingdom [2000] (ECtHR). Available at: http://hudoc.echr.coe.int/eng?i=001-58841 (Accessed: 12 September 2017).

Khelili v. Switzerland [2011] (ECtHR). Available at: http://hudoc.echr.coe.int/eng?i=002-345 (Accessed: 12 September 2017).

L.C.B v. The United Kingdom [1998] (ECtHR). Available at: https://hudoc.echr.coe.int/eng#{"itemid":["001-58176"]} (Accessed: 12 September 2017).

Lawless v. Ireland [1961] (ECtHR). Available at: https://hudoc.echr.coe.int/eng#{"itemid":["001-62076"]} (Accessed: 12 September 2017).

Leander v. Sweden [1987] (ECtHR). Available at: http://hudoc.echr.coe.int/eng?i=001-57519 (Accessed: 12 September 2017).

Loizidu v. Turkey [1996] (ECtHR). Available at: http://hudoc.echr.coe.int/eng?i=001-58007 (Accessed: 12 September 2017).

M.M. v The United Kingdom [2013] (ECtHR). Available at: http://hudoc.echr.coe.int/eng?i=001-114517 (Accessed: 12 September 2017).

Manitaras and Others v. Turkey [2008] (ECtHR). Available at: http://hudoc.echr.coe.int/eng?i=001-87232 (Accessed: 12 September 2017).

Morice v. France [2015] (ECtHR). Available at: http://hudoc.echr.coe.int/eng?i=001-154265 (Accessed: 12 September 2017).

Mouvement raelien Suisse v. Switzerland [2013] (ECtHR). Available at: http://hudoc.echr.coe.int/eng?i=001-112165 (Accessed: 12 September 2017).

Oneryildz v. Turkey [2004] (ECtHR). Available at: http://hudoc.echr.coe.int/eng?i=001-60510 (Accessed: 12 September 2017).

Osman v. The United Kingdom [1998] (ECtHR). Available at: http://hudoc.echr.coe.int/eng?i=001-58257 (Accessed: 12 September 2017).

Ouranio Toxo and Others v. Greece [2006] (ECtHR). Available at: http://hudoc.echr.coe.int/eng?i=001-70720 (Accessed: 12 September 2017).

Ozgur Gundem v. Turkey [2000] (ECtHR). Available at: http://hudoc.echr.coe.int/eng?i=001-58508 (Accessed: 12 September 2017).

Peck v. The United Kingdom [2003] (ECtHR). Available at: http://hudoc.echr.coe.int/eng?i=001-60898 (Accessed: 12 September 2017).

Rasmussen v. Denmark [1984] (ECtHR). Available at: http://hudoc.echr.coe.int/eng?i=001-57563 (Accessed: 12 September 2017).

Roman Zakharov v. Russia [2015] (ECtHR). Available at: http://hudoc.echr.coe.int/eng?i=001-159324 (Accessed: 12 September 2017).

Soering v. The United Kingdom [1989] (ECtHR). Available at: http://hudoc.echr.coe.int/eng?i=001-57619 (Accessed: 12 September 2017).

Stec and Others v. the United Kingdom [2006] (ECtHR). Available at: http://hudoc.echr.coe.int/eng?i=001-73198 (Accessed: 12 September 2017).

Uzun v. Germany [2010] (ECtHR). Available at: http://hudoc.echr.coe.int/eng?i=001-100293 (Accessed: 12 September 2017).

Von Hannover v. Germany (No. 2) [2012] (ECtHR). Available at: http://hudoc.echr.coe.int/eng?i=001-109029 (Accessed: 12 September 2017).

Weber and Saravia v. Germany [2006] (ECtHR). Available at: http://hudoc.echr.coe.int/eng?i=001-76586 (Accessed: 12 September 2017).

Zarb Adami v. Malta [2006] (ECtHR). Available at: http://hudoc.echr.coe.int/eng?i=001-75934 (Accessed: 12 September 2017).

International Treaties

Council of Europe, *European Convention for the Protection of Human Rights and Fundamental Freedoms, as amended by Protocols Nos. 11 and 14*, 4 November 1950, ETS 5. Available at: http://www.refworld.org/docid/3ae6b3b04.html (Accessed: 13 September 2017).

Council of Europe, *Protocol 1 to the European Convention for the Protection of Human Rights and Fundamental Freedoms*, 20 March 1952, ETS 9. Available at: http://www.refworld.org/docid/3ae6b38317.html (Accessed: 14 September 2017).

Council of Europe, *Protocol 12 to the European Convention on Human Rights and Fundamental Freedoms on the Prohibition of Discrimination*, 4 November 2000, ETS 177. Available at: http://www.refworld.org/docid/3ddd0cb44.html (Accessed: 14 September 2017).

UN General Assembly, *Convention on the Elimination of All Forms of Discrimination Against Women*, 18 December 1979, United Nations, Treaty Series, vol. 1249, p. 13. Available at: http://www.refworld.org/docid/3ae6b3970.html (Accessed: 13 September 2017).

UN General Assembly, *Convention on the Prevention and Punishment of the Crime of Genocide*, 9 December 1948, United Nations, Treaty Series, vol. 78, p. 277. Available at: http://www.refworld.org/docid/3ae6b3ac0.html (Accessed: 13 September 2017).

UN General Assembly, *Convention on the Rights of Persons with Disabilities*, 13 December 2006, A/RES/61/106, Annex I. Available at: http://www.refworld.org/docid/4680cd212.html (Accessed: 13 September 2017).

UN General Assembly, *International Convention on the Elimination of All Forms of Racial Discrimination*, 21 December 1965, United Nations, Treaty Series, vol. 660, p. 195. Available at: http://www.refworld.org/docid/3ae6b3940.html (Accessed: 13 September 2017).

UN General Assembly, *International Covenant on Civil and Political Rights*, 16 December 1966, United Nations, Treaty Series, vol. 999, p. 171. Available at: http://www.refworld.org/docid/3ae6b3aa0.html (Accessed: 13 September 2017).

UN General Assembly, *International Covenant on Economic, Social and Cultural Rights*, 16 December 1966, United Nations, Treaty Series, vol. 993, p. 3. Available at: http://www.refworld.org/docid/3ae6b36c0.html (Accessed: 13 September 2017).

UN General Assembly, *Universal Declaration of Human Rights*, 10 December 1948, 217 A (III). Available at: http://www.refworld.org/docid/3ae6b3712c.html (Accessed: 13 September 2017).

APPENDIX: SAMPLE INTERVIEW QUESTIONS

The following is a list of questions that were used in semi-structured interviews with emergency managers and technologists. The questions were not used in every interview. Interviews were semi-structured and dynamic, and in the case of technologists some were inapplicable for particular participants. Additional questions were posed for the purposes of follow-up or clarification. Nevertheless, the following list of questions is indicative of the lines of questioning followed.

Sample Questions: Emergency Managers

- Can you talk me though the process of the declaration of a major emergency?
- What kind of emergencies would occur most frequently?
- Can an emergency be declared an incident before it has actuated, if it's just an anticipated emergency?
- What are the [organisation's] primary sources of information for situational awareness during emergencies?
- What are the challenges to the acquisition of reliable and timely information?
- How do you go about confirming reports or information that you're receiving from the site of an emergency?
- When an emergency is declared do emergency managers and responders get any extraordinary powers to address the incident that they may not have during normal circumstances?
- Can you talk to me a little bit about the [organisation's] engagement with social media during incidents?
- What are the challenges to using social media effectively during emergency management?
- How do you believe that a system such as Slandail in particular will change information acquisition during emergency management and response?
- What are the challenges associated with implementing a system such as Slandail?
- Based on your understanding of the system, what added, or additional functionality, or capabilities do you think would benefit the [organisation] or emergency responders generally?

- Would the system be useful if it were adapted to be functional in other types of emergency or incident aside from natural disasters?
- Will the [organisation] have any need to store any data obtained from the system locally?
- How long do you think that you would need to hold on to any material that was saved?
- If a system such as Slandail were implemented, would it displace any current methods, or systems, or be supplementary?

Sample Questions: Technologists

- Can you describe in your own words the functions of [system component]?
- Can you describe, in the context of Slandail, the benefits of [system component] for emergency response and management?
- Which social media sites can it import information from? Are there any limitations there as to which sites it can use as sources?
- Does it preserve any text from social media documents?
- Is it theoretically possible to anonymise sensitive data in tweets like names and things like that?
- Does [system component] have the capacity to be useful outside of a natural disaster scenario? So can it yield useful information in the event of different types of emergencies such as man-made emergencies or etc.?
- Does [system component] have any features or tools that help to ensure the veracity or the truthfulness of the information in the source documents?
- Is there any element of machine learning in [system component]?
- To what extent is the training automated. Does it learn by itself or does it need a lot of human input?
- Does [system component] log user access?
- Where are the servers [that contain data collected by component] located?
- What data does it use to determine the location [of social media user]?
- Is there any possibility that tweets originating from outside the target region would be processed by the system?
- Can you talk me through what kind of information is presented to the end-user?