VALTIOTIETEELLINEN TIEDEKUNTA
STATSVETENSKAPLIGA FAKULTETEN
FACULTY OF SOCIAL SCIENCES
75

# COMPUTATIONAL TRANSFORMATION OF THE PUBLIC SPHERE

## Theories and Case Studies

Edited by SM Amadae

# Computational Transformation of the Public Sphere

## Theories and Case Studies

S M AMADAE, Editor

VALTIOTIETEELLINEN TIEDEKUNTA
STATSVETENSKAPLIGA FAKULTETEN
FACULTY OF SOCIAL SCIENCES
75

# AUTHORS

Saana H. Annala, Anna M. Bogdan, William E. Burden, Vanessa Garcia Torres, Juuli Hakulinen, Heli Hämäläinen, Helmi Hämäläinen, Heidi Hanhela, Anna Hattunen, Aleksander Heikkinen, Jooel Heinonen, Anni Helkovaara, Tiina M. Helojärvi, Tom Henriksson, Aino Hiltunen, Hannamari Hoikkala, Dongyang Huo, Sami Husa, Riikka T. Ilmonen, Kaarina Järventaus, Alina Juote, Anni Juusola, Lisa Kärnä, Lasse I. Keinonen, Judith Knebler, Roosa M. Kontiokari, Christa-Jemina Korhonen, Eetu Kukila, Laura Kuokkanen, Susanna Kupiainen, Oona Laine, Vilma E. Lappalainen, Mari Liukkonen, Noora Magd, Matilda Mahne, Alina Mäkynen, Juhani Mäntyranta, Annina Mattila, Iiris Meurman, Juho Mölsä, Juho Myyryläinen, Nuura H. Naboulsi, Emmi Nahi, Emilia Nieminen, Sakari Nuuttila, Eeva Nyman, Dominic O’Hagan, Waltteri Oinas, Leo H. Pahta, Riikka Pasanen, Eemeli Peltonen, Essi Pitkänen, Helmi Rantala, Matti Rantanen, Inka Reinola, Anniina Riikonen, Victoria Ristikangas, Mikaela Rydberg, Tommi Saarnio, Eero Salojärvi, Lumi Saukkonen, Roosa M. Savo, Lili Schatz, Elisa Seppänen, Marja Silvolahti, Sonja Siirtola, Adam Oliver Smith, Kaarlo Somerto, Nico Stockmann, Pietari Suomela, Eleanor Suovilla, Mio Tamakoshi, Anni M. Taskinen, Laura Trémouille, Elina Uutela, Veera Villikari, Vesa Vuolle, Altti Vuori, Johan Wahlsten, Julia Ylä-Outinen

UNIVERSITY OF HELSINKI
FACULTY OF SOCIAL SCIENCES

Published in Finland by the Faculty of Social Sciences, University of Helsinki
Unioninkatu 37 (P.O. Box 54)
University of Helsinki
Helsinki, 00014, Finland

Names: Amadae, SM, Editor

Title: Computational Transformation of the Public Sphere: Theories and Case Studies

Description: Helsinki, University of Helsinki, 2020 |
Includes bibliographic references.

Identifiers:

ISBN 978-951-51-6144-4 (paperback)
ISBN 978-951-51-6145-1 (PDF)
ISBN 978-951-51-6146-8 (EPUB)

Printed by Unigrafia,
Hiomotie 3, Helsinki, 00380, Finland

# PREFACE

*Computational Transformation of the Public Sphere* is the organic product of what turned out to be an effective collaboration between MA students and their professor in the Global Politics and Communication program in the Faculty of Social Sciences at the University of Helsinki, in the Fall of 2019. The course, Philosophy of Politics and Communication, is a gateway course into this MA program. As I had been eager to conduct research on the impact of new digital technologies and artificial intelligence (AI) on democratic governance, I saw this course as an opportunity to not only share, but also further develop my knowledge of this topic.

The curriculum, which featured Jürgen Habermas and John Rawls, in addition to the luminaries John von Neumann, Alan Turing, and Claude Shannon, was demanding. The course culminated with students exploring the impact of our contemporary information revolution on specific cases and applications. As well, students had the opportunity to develop their skills as researchers, and thus welcomed opportunities to generate their own insights and findings.

The course was designed so that a large component of work was devoted to research papers. In order to ensure that students had the time to delve into their topics, they worked in groups of five, with each team generating their own research questions and approaches. We vetted projects together, and each team member took on the responsibility for one component of their research.

The volume stands as a remarkable collection of essays because the papers are strikingly timely and innovative. The information age with its new technologies of big data, machine learning, and algorithmic governance is a dynamic field. Hence young scholars have an important role to play in both mastering past classics in the studies of politics and communication, and quickly coming to terms with the tremendous opportunities and challenges posed by the digital revolution. Our volume provides a marvelous overview of our current moment, wherein cybersecurity and algorithmic governance are as important as, and now inseparable from, nuclear weapons and democracy.

I am grateful to Vice Deans Juri Mykkänen and Juhana Aunesluoma for their intellectual encouragement and material support of this volume. I appreciate my conversations with Hanna Wass during which this project crystalized. Eeva Hagel and Kimmo Jokinen at Unigrafia Oy graciously printed the text. Veera Koskinen provided invaluable PR for this volume's celebratory launch commemorating the Faculty of Social Sciences 75th Anniversary.

SM Amadae, Editor

# INTRODUCTION

My spouse is giving a virtual philosophy talk at the Center for Philosophy of Social Science (TINT) at the University of Helsinki today. She will interact with her audience through a computer screen, and take questions from people who click on the “raise hand” icon, in the comfort of their own dwellings. This is the only way it is possible to give a lecture in the Spring of 2020, with everyone grounded, and tethered to their computers for the foreseeable future.

Lectures and classes are playing out on computer, tablet, and phone screens. Our education and scientific participation are now dependent on having the right software, the right device, and the right Internet connection. We need to make sure we are properly illuminated on the outside as well as on the inside, and we need to know when to mute our microphones and when to turn on our cameras. In the age of the pandemic, Zoom is the new Agora, where we have all been forced to gather as a result of a different kind of invasion.

Neither is there any kind of interpersonal communication, art, entertainment, education, commerce or yoga class taking place outside of our computers. While most of these arenas of our lives have had some technological elements for decades now, the sudden shift that took place in the beginning of 2020 is astonishing in its scope and impact on our lives. The interaction and interdependence between technology and society has never been so evident and all-encompassing.

The essays in this volume were written just a few months ago, and even when we started to edit them, life was “normal”. The issues they are concerned with, while recognized as important in principle, have often been considered more “academic” than “practical” in nature.  However, this is increasingly changing. Debates about fusing and perhaps replacing deliberative democracy with algorithmic governance, privacy with totalitarian surveillance, or the consumption of news with microtargeted propaganda messages, are heating up. Most of these trends are being sold to us under the guise of progress, hailed not only as desirable but often inevitable, over our heads and out of our hands. Whereas some of these changes may indeed be desirable, nothing we have the power to collectively decide on can ever be inevitable. Cynicism and fatalism are the biggest obstacles we face on the precarious path to continued human existence in a rapidly technologizing world.

The authors in this book agree that if we harness technology in the right ways, we can enhance the public sphere instead of allowing greed and ignorance to hollow it out. If we allow our inherent morality and wisdom to be part of the discussion around emerging AI technologies, we can all benefit from these innovations as a society, instead of having to write off some of our fellow humans. Using the digital tools at our disposal in the right way is also our only hope of being able to tackle global challenges such as security, climate change, or misogyny.

We start our journey of exploration in Finland, looking at ways to best utilize digital technology to fortify the country from the internal threat of the **erosion of the public sphere and democratic deliberation** (I.1), as well as from newly emerging external **cyberthreats**, by aptly addressing **securitization** (I.2). In both cases, we must walk a fine line between individual freedoms and the common good.

Next, we turn our attention to the new reality of politics around the world: refreshing, exciting, direct – and mendacious. We learn how the elevation of two reality TV actors to the presidency, one in the US and one in Ukraine, was made possible by hacking the electorate's collective brain to blur the line between reality and fiction, as well as to straight up spoon-feed us what we have come to call **alternative reality** (II.2,3). A comparative analysis of two referenda in the United Kingdom, the **Scottish Independence Referendum** and the **Brexit** vote, highlights the massive and alarming shift that took place in just the course of a couple of years, identifying the current political predicament as resulting, at least partially, from the surgical and cynical removal of "informed" from the dignified process of **informed decision-making** (II.1).

Being informed is necessary, but not always sufficient. The main challenge we need to overcome if we are going to neutralize **the existential threat posed by climate change** is to restore our trust in science, as well as our own ability to work together. In order to tackle the underlying problem, **the dilemma of cooperation**, we need not only a shared understanding of scientific reality, but also a common moral ground. In the battle for survival, everyone needs to take up their posts, from the officers of governments, through the correspondents of the media, down to each private citizen (III.2). Companies, increasingly as powerful as governments in many ways, must also do their part. If they are caught promoting themselves at the expense of the environment, and our common good, through unsubstantiated pledges of sustainability (**greenwashing**), they must be held to account, and forced to comply with the environmental values they claim to stand for (III.1).

In 1964, the Rev. Dr. Martin Luther King Jr. warned against the dangers of science without morality: "The richer we have become materially, the poorer we have become morally and spiritually," he said. "We have learned to fly in the air like birds and swim in the sea like fish, but we have not learned the simple art of living together as brothers."[1] Over half a century later, this is still as important a message as ever. As mobility has become increasingly more accessible, an examination of whether the emergence of ride-hailing applications offers **increased emancipation and freedom** for, in this case our sisters, is welcome (IV.1). With the arrival of a new kind of citizen to our ecosystem, the **human robot,** with whom (which?) we will also need to live together, more questions will need to be answered. In the case of Sophia, one of

[1] Nobel Lecture, December 11, 1964, https://www.nobelprize.org/prizes/peace/1964/king/lecture/

the first such robots, we certainly have more questions than answers: Is Sophia an it or a she? Is it alive? Should this humanoid have **robot rights**, and if so, should those be accompanied by obligations? Can robot rights be elevated to, or even exceed the extent of human rights? These questions will prompt us to reconsider not just our relationship to machines, their relationship to each other, but also the relationships among ourselves (IV.2).

Digital technology has the ability to affect the way we relate to each other in many different ways. It can be utilized to challenge our millennia old, ongoing battle against patriarchy, by offering us ways to connect, organize and make our voices heard, as we have seen in the case of the **#MeToo** movement (V.2). Technology can also be used to connect and organize by those left behind by the cruelty of neoliberalism, their unheard cries for help at the tormenting hands of some misguided expectation of masculinity **echoing** louder and louder in their **chambers**, until their frustration too often finds a violent outlet. Algorithms can help us find and rescue some of these victims who call themselves **involuntary celibates** but, ultimately, changes in the curriculum as well as in the public discourse to promote a healthy understanding of gender and sexuality will be necessary in order to eliminate gender-based violence (V.1).

Changes in legislation, as well as new laws will also be needed if we hope to keep up with the ways our consumption of news and political messages are changing through the introduction of new platforms of communication. **Fake news** is a clear and present danger to democracy, and yet it has been spreading unobstructed from the moment it was fanned into a global wildfire by **social media**, from the relatively harmless simmer of the tabloid media. We need to snap out of our helpless shock of watching institutionalized misinformation destroy everything in its path, from the credibility of science, through privacy, to the fairness of elections, and start making some tough decisions about how to put out the fire and rebuild what has been lost. This can be done through government regulation, by means of enlisting algorithmic tools, or by using a hybrid, **Nordic model** which promises to enhance citizens' ability to recognize and resist misinformation through **media literacy education** (VI.1). Attempts to regulate, or take advantage of, the crime scene have been made in accordance with prevailing norms and values. While most social media companies are chasing profit, the EU has initiated **GDPR** to salvage privacy, the UN has plans to use big data for humanitarian efforts, and the US government has enlisted private companies to secretly capture and **collect** as much **data on citizens** as possible. Meanwhile, China decided to skip the middleman, and made **surveillance** a matter of state policy (VI.2,3).

The question of how much control we are willing to hand over to **algorithms** in **governing our lives** is central throughout this volume. In the last two chapters the panda in the room is finally tackled head on. Can a political system that utilizes AI to realize **totalitarian surveillance** and control through an impersonal classification system of its population ever be legitimate? Can it possibly be desirable? Ethics, consent, opacity and accountability will need to be carefully considered before we can answer these questions. One thing is clear, however. At the same time that we turn our concerned and critical gaze towards China's social credit system, the West should not forget to put its own house in order with regard to its financial credit systems, and other processes increasingly outsourced to **"black box" algorithms** (VII.1,2).

The essays collected in this volume adhere to a disciplined, scientific approach of examining emerging digital technologies. They carefully weigh multiple considerations, very much including moral ones, present different sides of the argument, cite sources, and offer criticism as well as policy recommendations. Most authors conclude that more research, discussion and a wider consensus is needed before we can implement change, or offer the right reaction to changes that have already been implemented. Now, with an accelerated reliance on technology, and a multiverse of possibilities ranging from the dystopian to the utopian ahead of us, the application of this kind of measured and systematic approach is more imperative than ever.

We will, no doubt, defeat the current pandemic, and be able to attend lectures and visit museums again. But crises in our societies are ongoing, and new ones can be expected. The central question in all of these predicaments is whether a renewed sense of enlightenment, a resurgence of respect for knowledge, science, and morality is allowed to guide us through these dark tunnels, once again out into the light. When I read the essays in this volume, and consider the Faculty of Social Sciences at the University of Helsinki as a whole, I am hopeful that it will.

András Rátonyi, Managing Editor
April 16, 2020

# Part I

# Cyber Finland

# 1.1 The Impact of the Digital Revolution on Citizen-Governance Interaction in the Finnish Context

Marja Silvolahti, Eemeli Peltonen, Juho Myyryläinen, Oona Laine, Kaarina Järventaus
Faculty of Social Sciences, University of Helsinki

## Abstract

This research paper examines the impact of the digital revolution on citizens and governance. New technologies introduce tools and platforms that provide new forms of participation such as e-voting, initiatives and mass mobilization. The digital revolution and new technologies have improved citizens' political participation and engaged citizens in political decision-making. Alongside political inclusion, technology provides citizens with a platform for interaction with each other and political decision-makers, which can be seen as contributing to the deliberative development of society.

New technologies have improved the chance of citizens' participation, but the mere possibility of participation may not be sufficient to promote democracy and political participation. Engaging citizens in political decision-making may be challenging for “traditional reasons”: lack of political interest or knowledge. However, too much skepticism about citizen engagement is not entirely justified, as new technologies offer the potential to expand social capital.

*Keywords*: Electronic democracy, online deliberation, electronic voting, citizens' initiatives, crowdsourcing, participatory budgeting, citizen-governance interaction

# 1. Introduction

In this research paper we explore how the advent of new technologies in the 21st century have impacted the interaction between citizens and governance in Finland. By new technologies we refer to all the new tools and networking platforms brought about by the digital revolution. These tools and platforms under our investigation include social media, web-based programs and new means of electronic participation, for example e-voting.

We will narrow our focus to the Finnish experiences of new ways of communication in citizen-governance interaction. We expect to find case-based evidence about the implications of the digital revolution for the means of participation and conduct of governance. Our key research questions are:

- What implications has the digital revolution of the 21st century had on citizen-governance interaction in the Finnish context?
- How have new tools and technologies changed the ways of conducting governance?
- How could the Internet be made more friendly for democratic deliberation?
- What problems does the digital revolution pose for participation, governance and respectful deliberation as well as a properly functioning public sphere?

We decided to begin our research paper by defining some of the key concepts of our research, in order to have a common base for our individual contributions. Next, we move on to explore five specific cases of citizen-governance interaction in the digital era of the 21st century. In the first chapter, we explore how the Internet could be made more friendly to democratic deliberation. The second chapter addresses the question of e-democracy through a crucial practical aspect of it, namely electronic voting. The third case study looks at citizen initiatives: do they really matter and what motivates citizens to try to achieve their political goals through them? Our two final chapters explore cases of crowdsourcing in policy-making and participatory budgeting in the Finnish context.

# 2. Key concepts

### Digital revolution

Despite the widespread use of this concept, it is not always defined specifically in contemporary everyday use. The digital revolution is all about broad technological *changes* in politics, culture, economics and business. As a result of these changes digital technology will take a more prominent and central place in our everyday lives. No aspect of human life, communication or business will be completely separated from the digital revolution, which continues to expand (Meyer 2016).

The digital revolution includes the *"rise of digital platforms, cutting edge forms of automation and Big Data"* (Spence 2019). Although the concept of the digital revolution is already vastly used, we have not yet seen all of its implications.  Thus, we argue that the digital revolution will change our politics, economics and business during the time period of the 21st century as we proceed into the future.

Luciano Floridi goes even further to describe how the digital revolution affects our everyday lives. He argues that the digital revolution will also change our social selves and the way we see ourselves. The digital revolution and ICT have put forward new concerns of privacy, openness and transparency. (Floridi 2014) These examples from Floridi's writings exemplify how profoundly the digital revolution will shape our world.

Henning Meyer argues that *"there is a general lack of structured analysis of the ways in which technological progress translates into real life"* (Meyer 2016). Thus, our research paper aims to capture one dimension of this large and complex transformation by examining the impact of the digital revolution on citizen-governance interaction in the Finnish context during recent years.

We expect to find evidence for growing interconnectedness between citizens and government. We also hope to build a general view on the current situation of how new digital tools are used by Finnish governance. Can the digital revolution bring citizens closer to those who aspire to rule them? Have the challenges in citizen-governance interaction changed with digital revolution, or are they more deep-rooted by nature?

## Public sphere and public opinion

The theoretical point of view in this research paper is the thinking by Jürgen Habermas regarding the public sphere. Habermas defines this useful term in his article "The Public Sphere: An Encyclopedia Article" (1964) as follows:

> By the "public sphere" we mean first of all a realm of our social life in which something approaching *public opinion* can be formed. *Access is guaranteed to all citizens.* A portion of the public sphere comes into being in every conversation in which private citizens assemble to form a public body…Citizens behave as a public body when they confer in an unrestricted fashion—that is, with the guarantee of freedom of assembly and association and the freedom to express and publish their opinions—about *matters of general interest.* (Habermas 1964, p. 50, italics added)

The formation and manifestation of public opinion is relevant in all the themes we address throughout this paper. A democratic system of governance requires these two processes in order to represent the political will of the citizens that it represents. Accurate representation is also the goal of the equal and unrestricted access that Habermas mentions.

The digital revolution has had an undeniable impact on the public sphere and the forming of public opinion. This development has brought to life new platforms and forums, annihilated the significance of geographical distance in communication between individuals, and made it easy for people to create their own publicities around what matters to them. It allows decision-makers and the governance in general to collect data about peoples' opinions and preferences and to hear the voice of the general public more easily than before. In these ways, the digital revolution shortens the distance both between citizens and between citizens and governance. On the other hand, much of the discussion that takes place online does not necessarily count as quality deliberation, and the opinions that come across from discussions are not necessarily manifestations of public opinion as opposed to the voice of a few loud actors.

### Governance

Andrew Heywood defines governance as "…a broader term than government. Although it still has no settled or agreed definition, it refers, in its widest sense, to the various ways through which social life is coordinated. Governments can therefore be seen as one of the institutions involved in governance" (Heywood 2007). In this research paper we use this definition. We prefer to use a concept that is reasonably wide to capture as many aspects of governance as possible in our Finnish context.

Heywood continues to define three principal modes of governance: markets, hierarchies and networks (Heywood 2007). Our research will be focused on the latter two of these modes: hierarchies and networks. We are especially interested in tools of governance that are used to amalgamate and coordinate citizens' preferences to policy-outputs and policies. In Habermas' words these tools advance participation in the forming of public opinion, and inform the governance of it. For example, citizen initiatives and e-voting are such tools.

We see four distinct values central to governance: participation, communication, pluralism and accountability. Governments have many tools for conducting governance which fulfill these values. Democratic governance has to keep up with the changing technological challenges and opportunities, and we wish to shed some light on how the Finnish government has succeeded in this during the recent years.

## 3. How to make the Internet more friendly for democratic deliberation?

The Internet has become an integral part of the public sphere, where people can debate about current political and societal issues. Despite its many possibilities, the Internet is not always an ideal environment for democratic deliberation. For example, anonymity, hate-speech and bubbles can weaken the Internet as a platform for good and respectful deliberation between individuals. In this research paper I investigate how the Internet could be made friendlier to democratic deliberation.

## Deliberative democracy and the Internet

André Bächtiger, John S. Dryzek, Jane Mansbridge and Mark E. Warren define deliberative democracy to mean "mutual communication that involves weighing and reflecting on preferences, values, and interests regarding matters of common concern" (Bächtiger et al. 2018, p. 2). I use this definition of deliberative democracy here because of its clarity and simplicity, compared to other definitions presented in academic research.

Democratic deliberation refers to all forms of deliberation that fulfill democratic ideals. These ideal can also be used as standards of good deliberation. According to second generation theorists of deliberative democracy, standards of good deliberation include mutual respect, absence of power, inclusion, aim at consensus, publicity and accountability (Bächtinger et al. 2018, p. 4).

Theorists of deliberative democracy highlight how democratic deliberation can occur in many distinct sites. For example, formal institutions of government and civil society are often mentioned as sites for democratic deliberation (Bächtiger et al. 2018, p. 11). The Internet can also be seen as a site for democratic deliberation, because of its rapid rise as an important part of today's public sphere. When considering Jürgen Habermas' definition of the public sphere (Habermas 1991, p. 30), I locate the Internet's existence between the private realm and the sphere of public authority, much like the pre-Intenet public sphere in the political realm which existed between private households and state authority.

Democratic deliberation can occur on the Internet in many forms. Blogs, open-access websites and comment sections are places where everyone can easily discuss and deliberate about current issues. Popular social media platforms such as Facebook, Twitter and Instagram also offer many opportunities for democratic deliberation between citizens, without limitations of time or space.

## Problems of democratic deliberation on the Internet

I measure the quality of democratic deliberation on the Internet by using standards of good deliberation described earlier. The Internet differs from other sites of democratic deliberation in many ways. These features complicate the fulfillment of the standards for good deliberation.

One of the most striking qualities of the Internet is the absence of editors on many communication platforms. Traditional newspapers and magazines have editors who control what opinions and comments are published in the paper. However, many websites do not have any editors controlling the flow of messages and comments between individuals. This absence of moderation threatens the standards of good deliberation, because no one monitors or controls these discussions in order to ensure the quality of their content.

Absence of moderation can lead to a proliferation of hate-speech and disrespectful debate. There is a lot of evidence about this dark side of communication on the Internet. Recently it has been found that more than half of Americans have experienced harassment, hateful speech, physical threats and bigotry when using the Internet (USA Today, 2019). Hate-speech violates many standards of good deliberation, for example mutual respect, inclusion and the absence of coercive power.

It has also been found that the growing use of new social media platforms (such as Facebook, Instagram and Twitter) has increased the amount of fake news and disinformation (Martens et al. 2018, p. 8-10). Fake news poses a clear threat to good standards of deliberation.

Another distinct quality of Internet-based deliberation is anonymity. Many communication platforms like 4chan and reddit offer posters possibilities for deliberation without revealing their real name. Fake accounts are also easy to create on Facebook and Twitter. When compared to traditional media, there is usually some kind of restriction on anonymity (although nicknames can be used after a binding registration, for example in comment sections for newspaper articles on the Internet).

The Internet, and especially social media, can also create social bubbles, filtering of information and group polarization, which endanger good deliberation (Sunstein 2017, p. 59-97). Previous research on Internet-based communication supports the conclusion of the Internet fostering communication with already like-minded people (Sunstein 2017, p. 76-77). Thus, large online groups can spend years communicating with like-minded citizens without ever hearing or reading contrary views.

This phenomenon increases the risk of group polarization in society. Citizens communicating mainly inside their own social bubble become increasingly isolated from other groups. This is troubling because when various cross sections of groups are communicating *"society will hear a far wider range of views"* (Sunstein 2017, 86). Polarization between groups can also increase anxiety and suspicions about people communicating in a different social bubble or group.

### How to make the Internet more friendly for democratic deliberation?

After discussing problems of Internet-based communication, I will now move forward to offer solutions for making the Internet friendlier for democratic deliberation. First, I will underline some basic elements of the Internet that help advance democratic deliberation.

The Internet is clearly more accessible for citizens than traditional media platforms. Not everyone can get their response published on the pages of *Newsweek* or the *New York Times*, for example. The Internet provides a platform of publication for almost any author with any kind of content imaginable.

The Internet is also a highly visible site for deliberation. Comment sections and open Facebook groups are good examples of highly public sites for democratic deliberation. Anyone can read discussions taking place in open Facebook groups. Usually only registration is required before you can post your own comment or response to a news article on a newspaper's public discussion forum.

Despite these advantages, I cited some problems relating to Internet-based communication above. These problems underline the need for more inclusive, open and moderated discussions. Cass R. Sunstein has proposed seven solutions for making the Internet more friendly for democratic deliberation: deliberative domains, disclosure of relevant conduct, voluntary self-regulation, economic subsidies for public networks, must-carry policies for media, creative use of links to draw people's attention to multiple views and opposing viewpoints (or serendipity) buttons (Sunstein 2017, p. 215; Economist 2017). Sunstein's proposals are presented and further described in the table below.

**Cass R. Sunstein's proposal for increasing democratic deliberation on the Internet (Sunstein 2017)**

| **Proposal** | **Description of the proposal** |
|---|---|
| Deliberative domains | Moderated (or edited) platforms for democratic deliberation between individual citizens. |
| Disclosure of relevant conduct | Policies aimed to encourage media to disclose relevant information about their content (for example, giving information about the suitability of programming on television). |
| Voluntary self-regulation | Voluntary self-regulation is about media companies regulating themselves, for example by providing a wide range of views for the public. |
| Economic subsidies for public networks | Public funding for media companies which aims to avoid polarization and consumerism of the news media. |
| Must-carry policies for media | Legislation, which requires media companies to provide the public with specific relevant information, news and programs (for example, about political debates, elections and democratic principles). |
| Creative use of links | Offering readers opposing viewpoints via links to different articles. |
| Opposing viewpoint buttons | Buttons that provide opposing viewpoints for users who are interested in them (for example, after reading a news article about a specific topic). |

Sunstein's deliberative domains are platforms where discussion is moderated. These platforms offer spaces where citizens can meet and deliberate about different topics. The aim of these platforms is to foster better understanding, learning and citizen engagement (Sunstein 2017, p. 216-217). Deliberative domains can be seen as somewhat naïve because, for example, citizens do not have any incentive to switch from Facebook to deliberative domains if they are satisfied and familiar with communication on Facebook.

A creative use of links and opposing viewpoint buttons can be seen as a more realistic solutions for problems relating to deliberation on the Internet. According to

Sunstein, newspapers and digital platforms should also offer readers articles that contain different viewpoints on the topic they have previously read about (Sunstein 2017, p. 229). This creative use of links could expose readers to a diverse range of information through other readers' stances and beliefs. Sunstein also proposes the introduction of opposing viewpoint buttons, which would offer readers *"opposing viewpoints by default, subject to the right to opt out"* (Sunstein 2017, p. 232-233). For example, these buttons could be included in web-based articles of a specific newspaper's digital edition or all of a newspaper's website articles.

### Concluding thoughts

As I have stated above, the Internet has both advantages and disadvantages for democratic deliberation. I have listed a few proposals from existing research for making the Internet friendlier for democratic deliberation. I view these proposals with skepticism because they would require a complete reform of the Internet. Since no one owns or controls the Internet, these proposals would be extremely difficult to implement. Thus, it is up to governments, organizations, companies and individuals to come together and try to carry out solutions that fulfill standards of good deliberation on Internet-based platforms. These solutions do not come easily and require extensive co-operation between different actors.

## 4. Electronic voting in Finland – many attempts, little success

Moving on from the process of deliberation to finding ways to organize society in accordance with the results of deliberation, voting and elections are the focus of this chapter.  A manifestation of "something approaching public opinion" (Habermas 1962; p. 1974) created in the public sphere is the goal of organizing elections. Democracy requires a way of getting to know how people think and what they want. In representative democracies such as Finland, elections aim at selecting a group of people that would share the thoughts of those who voted for them and—at least ideally—represent the citizens as accurately as possible. In the Habermasian sense, the formation of public opinion happens in the deliberation prior to the actual election, and the election result provides a concrete outline of it.

Voting is an encounter between the private and the public spheres. It is an act of an individual, as each citizen enters the voting booth alone. The secrecy of the ballot is an institutionalized principle to guarantee that all citizens can indeed decide for themselves without having to worry about social consequences or their vote going public. In Finland this principle is secured in the constitution (731/1999, 25 §). On the other hand, much of what happens prior to and after the casting of the vote is very much public. In elections the governance and the citizen come together in a concrete way.

**Could electronic voting bring people to the ballots?**

Turnout in Finnish elections has been decreasing since the 1970s. This has been seen as a signal of growing political disinterestedness, and even the proverbial “crisis of democracy”. In the 2019 parliamentary elections the turnout was 72,1 % (Statistics Finland). Research shows that the level of political participation varies across Finland, and turnout percentages are the lowest among groups that are in socio-economically weaker positions in society (Wass & Borg, 2016). An often-proposed solution for this participation challenge is electronic voting: maybe if those who do not go and vote at their designated voting place would cast their vote if they could do it where they wish to, using their computers or smartphones.

In recent decades electronic voting has been considered by working committees under several different cabinets. The most recent case was a working group requested in 2016 when Juha Sipilä, prime minister and the head of the cabinet bearing his name, claimed that Finland would be moving to electronic voting in the future, with the traditional paper method of voting still continuing to exist as an option. However, the working group came to the conclusion that the risks of online voting outweigh its benefits. Several issues were identified, namely the reconciliation of verifiability and election secrecy (the data of the voter would have to be stored alongside with the vote so that it could be later verified, but this would be illegal and compromise the secrecy of the ballot), manipulation of election results, breaching of election secrecy, and external interference through denial-of-service attacks. The biggest concern named by the working group was the loss of public confidence, which could easily be caused by spreading disinformation and rumors.

Another counterargument to an electronic voting system that the working group identifies is that it does not seem to solve the problem of decreasing voter turnout. It refers to several Nordic and Canadian research projects which have found that electronic voting does not increase turnout (Bergh & Christensen 2012; Segaard, Bock, Baldersheim & Saglie 2012; Bochsler 2010) and that the people who do vote electronically are ones that would vote anyway, regardless of the method (Goodman 2014). The only research this report refers to that gives a positive estimation of potential increase in turnout assesses this increase to be a rather modest percentage, around 2-3% (Vassil, Kristjan, Weber, Till 2011).

Finnish citizens’ attitudes towards electronic voting have not been representatively surveyed. The closest things to such a survey are two municipal level democracy ARTTU surveys conducted in 2008 and 2011, where people living in a cluster of Finnish municipalities were asked to agree or disagree with the claim, “People should be allowed to vote via the Internet in municipal elections”. A more recent Special Eurobarometer Survey requested by the European Commission (2018) mapped European Union citizens’ thoughts on new voting methods, and their concerns are in line with these challenges identified by the Finnish working group.

Citizens were asked to imagine that they were able to vote electronically, online or by post, and then to name their possible concerns about voting using these methods. The most cited concern in the survey was the potential for fraud or cyberattack: 68% said they were concerned, one third (33%) were very concerned. The idea that these systems posed difficulties to some segments of the population, such as people with disabilities or older people, was a concern for almost two thirds (65%) of the respondents. More than half were also concerned about voters being influenced by third parties (56%), and about the secrecy of the ballot (55%), with 23% and 24% very concerned, respectively.

In the 2008 municipal elections, electronic voting was tested in three municipalities. The system did not work correctly, and in 2009 the Supreme Administrative Court of Finland decided that voting was to be reorganized in these three municipalities. In addition to technical issues, the instructions provided were unclear and insufficient. No tests have been conducted since. However, the results of the 2008 test are valuable and point out that a functional electronic voting system requires more than just reliable technology—the voter needs to be well instructed and there cannot be any ambiguity in the communication.

### Electoral term 2019-2023: no plans for e-voting

In 2019, the recently appointed Ministers of the Interior (Maria Ohisalo, Green League) and Justice (Anna-Maja Henriksson, Swedish People's Party) both stated that for now, there is no need to consider updating the Finnish voting system. In a news article by Yle their opinions on the issue were very like-minded: the present system may be a bit "old-fashioned", but it is both secure and functional, and alternative methods presented to this day are too risky. With these statements and the pessimistic conclusions in the report requested by the Juha Sipilä cabinet, it seems unlikely that Finland would be taking any steps towards an electronic voting system in the upcoming years.

The Nordic countries would have many advantages if they wanted to be e-voting frontrunners. These countries are known to be some of the most stable democracies in the world, with little corruption and next to no past cases of electoral fraud. They are countries of high educational level, and the welfare state model allows willing states to invest in completely tax-paid development and research projects. One can think that another advantage are the relatively small populations in the Nordic countries, with a population of 4-6 million in most of them, and only 10 million in Sweden. Adopting a new large-scale system on the state level seems intuitively easier and more flexible in less populated countries. However, whether electronic voting will be advanced in the future is as much a question of trust as it is of technological development.

## 5. Civic initiatives – do they really matter, especially in the digital era?

In addition to e-voting, there are also other prominent ways to increase citizens' role in a deliberative democracy. Democracy researcher Rolf Büchi has said that direct democracy is a subtle process and its relevant elements are start, public conversation, decision-making and implementation (Büchi 2011, 107). Perhaps one of the biggest phenomena in recent years for that "starting point" has been the civic initiative. When we talk about civic initiatives, we mean procedures that "allow citizens to bring new issues to the political agenda through collective action, that is, through collecting a certain number of signatures in support of a policy proposal" (Schiller & Setälä 2012, p. 1).

### Kansalaisaloite.fi

In Finland, citizens' initiative is a tool for direct democracy which enables a minimum of 50,000 Finnish citizens of voting age to submit an initiative to the Parliament of Finland to enact an act (Väestörekisterikeskus). The initiative must include a bill or a proposal to start drafting legislation and the reasons for the proposal, and it must also apply to a matter that can be enacted by law. The development of digital technology, the Internet and social media has really speeded up the meaning of civic initiatives. The Ministry of Justice in Finland has set up an online system to collect statements of support; namely, Kansalaisaloite.fi where anyone can open an initiative and collect signatures. It is also possible for citizens to organize a municipal citizens' initiative (Kuntalaisaloite.fi) or simply editorialize some societal or political question (Otakantaa.fi).

It is obvious that when collecting civic initiatives online, space does not matter anymore, and time becomes more flexible, as well. People can sign civic initiatives anywhere and anytime; and by using social media channels such as Facebook, civic initiatives can spread rapidly and far. It is possible to speak about "mediated relations" (Grossi 2011, p. 4), which are unbound by time and space and which concern both people and their relations with organisations, institutions, places, goods, and objects. So, we can also see such a small concept, online initiatives, as a notable means to building and maintaining a lively public sphere where people and their civic society engage politicians and the state, in other words, governance. When assembling virtually for some topic, initiatives require direct deliberation of the Finnish Parliament. I interpret this as a converging relationship between private sorrows or hopes and public authority and formal politics. Initiatives create a picture of citizens' lives when they handle such themes as maternity law, free second-degree education or euthanasia (Kansalaisaloite.fi 2019). Browsing the website provides a remarkable insight into what citizens are worried about, as all civic initiatives are displayed in the same place.

But the whole picture is not that simple. We cannot claim that signing civic initiatives is an unequivocal solution for challenges of today's public sphere and deliberative democracy. Even though it could be claimed that civic initiatives aggregate citizens together and create a temporary community which forms at least a part of public opinion or public thoughts, initiatives could be claimed to emphasize only “liberal-individualist digital democracy” (Dahlberg, 2011). Digital media is understood here as “enabling individuals to gain the information they need to examine competing political positions and problems” and also providing them with the means for the registration, and subsequent aggregation (Dahlberg 2011, p. 861). Instead, civic initiatives do not formulate deliberatively constituted consensus, that is to say, rational public opinion, or make people argue, inform, reflect or publicize (Dahlberg 2011, p. 11).

So, civic initiatives still keep people and their opinions somewhat apart from each other, and they are purely individual choices and not vehicles for deliberation. We are not so communicative when we sign initiatives. We are able to surf the Internet and skip all the initiatives or just choose one which appeals to us. The facility to sign or not to sign concerns also the launching of civic initiatives. But launching or signing does not mean that something will happen. At the moment there are 69 ongoing and 955 completed civic initiatives on Kansalaisaloite.fi, but only 28 civic initiatives have been delivered to the Finnish Parliament. Also, we must remember that when an initiative is referred to parliament, representatives still have a right to do whatever they decide to do with it. It is conventional that civic initiatives do not pass as such, which will probably be the case of Suostumus2018 which demands a new rape law with a clear mention of consent (Yle 2019).

**Such a marginal vehicle?**

It could be claimed that the number of initiatives reflects the notion of a consumer-citizen. This kind of a citizen has so many choices that it is not important to even think deeply about them; so, you can sign civic initiatives even if you are not sufficiently informed about it, simply because you can. Giorgio Grossi uses the term “audience democracy” as a central concept. According to Grossi, we have moved from the space for discussion and formation of public opinion to a mere area of projective and symbolic identification typical of the ”society of the spectacle” and a new social environment which only favors the personalisation of choices and walks of life (Grossi 2011, p. 7).

This also causes the loosening of unsatisfactory cooperative and solidaristic ties (Grossi 2011, p. 6). This change in society and political life will deepen even more as the ongoing digital revolution continues. It is natural to differentiate one from the other and construct our own social, political and cultural bubbles when the Internet and especially social media are so broad, even boundless. The website Kansalaisaloite.fi might be precisely defined because the Ministry of Justice in Finland manages it, but there are many more areas and ways of trying to connect citizens, for instance the website Adressit.fi.

On the other hand, we need to remember that civic initiatives are a marginal vehicle for change, and people often do not give them much weight. And, as we have noted before, the possibilities inherent in online devices do not guarantee that people who have never participated in politics and deliberation would suddenly change their attitudes. So, it is worth mentioning that the digital revolution does not automatically cause any kind of revolution of citizen-thinking; those who were and had been active and committed will continue to do so in the digital era. People have tools, but also free will to decide whether to use them or not.

Hence, civic initiatives online cannot be scrutinized from one perspective. We do claim that they serve as a significant connection between citizens and governance, and a possibility to participate and reflect on a plurality of opinions. Kansalaisaloite.fi is a channel for citizens to have an effect on governmental decisions and debates, and they bring private and public questions much closer together. The future will show if civic initiatives somehow extend and get a bigger value both among administrations and citizens, and if they lead to a different and more solid kind of citizen-governance-relationship. Grossi (2011) also uses the terms “transnational individualised societies” and “global village” in considering today’s public sphere. The public sphere and public opinion can indeed be universal by means of global civic initiatives, but first they need to get their established place in a national social sphere.

## 6. Can crowdsourcing in policymaking foster democratic deliberation?

The digital revolution, especially networked online technology has made it possible for organisations and individuals to turn to a wider community of people to resolve problems and create new products. Initially a business concept, in recent years, the practice of crowdsourcing has been gaining ground also as a valuable tool in policymaking processes.

### Definitions and different contexts of use

The concept and definition of crowdsourcing was first presented by the editor of Wired Magazine, Jeff Howe, in 2006: "Simply defined, crowdsourcing represents the act of a company or institution taking a function once performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an open call. This can take the form of peer-production (when the job is performed collaboratively), but is also often undertaken by sole individuals. The crucial prerequisite is the use of the open call format and the large network of potential laborers." The new business model had early successes in creative and design industries as well as in corporate scientific research and development. (Brabham 2008, p. 76-79)

Brabham (2008) defines crowdsourcing as "a strategic model to attract an interested, motivated crowd of individuals capable of providing solutions superior in quality and quantity to those that even traditional forms of business can." He quotes Surowiecki who argues that "under the right circumstances, groups are remarkably intelligent, and are often smarter than the smartest people in them" (Brabham 2008, p. 79.). According to Aitamurto and Chen (2017), crowdsourcing "means an open call for anybody to participate in an online task". Outside business, crowdsourcing is used in very different fields like journalism, citizen science and crisis management (Aitamurto and Chen 2017, p. 3).

Crowdsourcing has been used in several countries in processes related to legislation or policy strategies: in constitutional processes in Iceland and Chile, in legislative or legal reforms in Finland and in Brazil, in strategy reforms by various US federal agencies, and on the municipal level in numerous urban planning projects. In policymaking, crowdsourcing can be used in different phases from problem identification and data gathering to developing proposals and consultation and to drafting policies, as well as implementing and evaluating decisions. In decision-making it is not applied very often, because in representative democracies decisions are made by elected bodies (Aitamurto & Chen 2017, p. 1-4).

## The Finnish case

The Finnish Ministry of Environment and the Committee for the future of the Finnish Parliament decided for the first time to use crowdsourcing in a law reform in 2013. The law in question was the off-road traffic law. Off-road traffic in Finland consists mostly of snowmobiles and all-terrain.  Off-road traffic had increased substantially and this created various controversies. Crowdsourcing was used in the research and drafting phase of the new legislation. Usually when laws are prepared, civil servants have direct contacts with different interest groups and with an expert committee with different stakeholders. This time, the Ministry also wanted to use crowdsourcing, aiming in its own words to “search for knowledge and ideas from the crowd, enhance people’s understanding of the law, and attempt to increase the perception of the policy’s legitimacy” (Aitamurto & Landemore 2016, p. 177-178).

In 2013, an online platform was opened where citizens could propose ideas, vote others´ ideas up or down, and comment. There were two crowdsourcing phases that generated about 500 ideas and 4,000 comments as well as 24,000 up or down votes from about 700 users altogether. The first phase was dedicated to problem mapping and the second phase aimed to generate and evaluate problem solving ideas.  The authors of the research article were active in designing the crowdsourcing platform. The platform was open to everybody, but to participate actively, one had to register with a verifiable e-mail address. (Aitamurto & Landemore 2016, p. 178-179)

As a practice, crowdsourcing enhances democratic value in several ways. It increases transparency both among peers and between the crowd and crowdsourcers, who are often policymakers. It informs citizens when the projects are still in their planning phases. It also increases inclusiveness, as it invites a large number of citizens to participate in policymaking. (Aitamurto and Chen 2017, p. 1-12)

Amoretti distinguishes between four types of e-democracy: consultative, participative, deliberative and administrative e-democracies (Amoretti 2006, p. 11-13). Crowdsourcing can be either consultative, participative or deliberative. In the Finnish case, the process was designed to be participatory, and that goal was reached, as 700 citizens participated actively. The question is whether it was also a deliberative process. Was the quality of discursive process emphasized, as well as rational reasoning?

The classical standards of good deliberation are: respect, absence of power, equality, reasons, aim at consensus, common good orientation, publicity, accountability and sincerity (Bächtiger & al. 2018, p. 4). In the Finnish case, many of these standards were well actualized and some others to a certain degree, even if the process was not designed to be a deliberative process.

Aitamurto and Landemore argue that the participants could act freely and that they were equal. On the other hand, they admit that the tone was not always respectful (Aitamurto & Landemore 2016, p. 186-188). Participants exchanged arguments in a dialogical manner. It was possible to distinguish arguments, counter-arguments, examples, counter-examples, conceptual distinctions, new propositions and use of evaluative criteria (Aitamurto & Landemore 2016, p. 182-186). These observations indicate that reasoning did happen, that there was a common good orientation and that at least many participants aimed at common understanding, if not at perfect consensus. On the other hand, the common goal was not to find a consensus but to communicate a plurality of aspects and experiences related to off-road traffic to policy-makers. The criterion of publicity was fulfilled, as the debate happened openly on a public platform. When it comes to participants, a minimum standard of accountability was achieved by the use of e-mail addresses. As for policy-makers, the crowdsourcing process increased their accountability, since the process of lawmaking was made transparent and because it is more difficult to ignore citizens´ views once they have been asked for and published.

It is important to remember that the crowdsourcing process cannot be seen as conveying “public opinion”. Technically, the material is a self-selected sample and it is not representative (Aitamurto and Chen 2017, p. 5-6; Brabham, p. 86). Brabham sees a risk of strengthening already extant hegemonies and suggests that we should keep a “constant eye on who is missing from the crowd” (Brabham, p. 86-87). Compared to traditional administrative and political approaches, crowdsourcing adds inclusiveness to the process. Still, the ideal of inclusiveness could probably be pursued even better than was done in the Finnish case.

## Concluding thoughts

To conclude this section, it can be said that an analysis of the Finnish crowdsourced law reform proves that crowdsourcing can include a relatively high degree of democratic deliberation. This seemed to be the case even though the process was not designed with that goal in mind. An important factor was certainly the deliberative domain (platform) skillfully founded and moderated by civil servants and media specialists. The process clearly increased both communication and participation on the citizen level and favored the expression of a plurality of views. It also increased inclusiveness, but did not resolve all the problems related to it.

## 7. Has participatory budgeting empowered citizens?

It can be said that democracy cannot only be taught to citizens on a theoretical level, but that the political and economic skills it requires can be learned by taking different actions in different arenas. Participatory budgeting (PB) is an example of one such trend. Participatory budgeting represents citizen involvement in public decision-making. This encompasses both democratic and economic innovation at the same time: PB can be said to give people "real power over real money" (Shah 2007, p. 45-47). The basic idea behind PB is to enable citizens to influence the use of public money in their own region when deciding on a budget and voting on viable ideas. Hence PB is meant to lead citizen participation and pluralism of economic power. The concept behind PB is the concept of participatory democracy and can be theoretically seen to follow the ideal of a deliberative concept of democracy: enabling citizens to debate ideas to be implemented: the best and most feasible idea is selected through social debate (Godwin 2018, p. 5-6). In public debate, citizens have the opportunity to define and prioritize the use of public wealth. As a complementary means of citizens' participation, PB represents the decentralization of power, which is associated with citizens' ability to make decisions.

Participatory budgeting can be characterized as a process of democratic decision-making and societal debate that offers the public sector the opportunity to encourage citizens to participate in political decision-making by accelerating democracy projects. The starting point for PB is therefore very grassroots, as citizens are the best experts on their needs in their daily lives, which can be seen to enhance the use of public funds and direct them to their most beneficial purposes (Godwin 2018, p. 12). Public sector funds consist of taxes at both the state and municipal levels, so it is reasonable to argue that it is financially fair to give citizens the opportunity to decide on the allocation of budgeted expenditure, which requires coherent communication between individuals and the public sector.

PB can be implemented by allocating part of the municipal budget for participatory budgeting. This will bring economic decision-making closer to the people and make it more likely that they will become more actively involved in the implementation of new ideas, which can be seen as reducing political inactivity and reticence. The idea of participatory budgeting also includes the concept of the budget being open, which means that it must be accessible to everyone, as well as easily understood. The budget can be visualized to improve comprehensibility which at the same time lowers the threshold for citizens to become familiar with the various stages of participatory budgeting (Shah 2007, p. 39).

In practice, PB goes through five stages: process design, brainstorming ideas, develop proposals, voting and project funding. Through its phases, PB seems, at the theoretical level, to be a complete means of controlling public funds and a stepping stone towards political inclusion and social justice. On the other hand, there are also problems that can be identified, especially at the voting stage: to represent a workable fund allocation tool and to be legitimate, it should involve enough citizens with sufficient political knowledge (Godwin 2018, p. 10-11).

**Participatory budgeting in Helsinki**

In the autumn of 2019, participatory budgeting is a politically relevant topic in Helsinki, as it is in the process of being implemented. In Helsinki, PB is practiced on two different levels: “OmaStadi” service and the “RuutiBudjetti” (Omastadi 2019), which is aimed specifically at young people. Implementing participatory budgeting in Helsinki is a relatively new social innovation, implemented in two phases. First, citizens—regardless of age or place of residence—were allowed to make proposals for the use of public funds. In the second phase, feasible proposals are voted on (Hel.fi 2019).

The implementation of PB represents the deliberative aspect of digital democracy, offering citizens the ability to decide on the resources to be allocated through electronic means in an online environment. On the other hand, participatory budgeting also approaches autonomous Marxism as it provides citizens with a path of political participation through cooperation between the individual and the public sector in the context of resource allocation. According to Dahlberg (2011), autonomous Marxism seeks a radical change that enables individuals to increase their influence through decentralization. In other words, Helsinki can be seen as contributing to the realization of local democracy through a participatory budgeting project that gives citizens budgetary authority while narrowing the gap between political decision-makers and citizens through the use of the online environment. It is this online environment that creates a space for public decision-making in which decentralized decision-making can take place, with the participation of people other than those who have obtained a political mandate.

In the second phase of participatory budgeting, Helsinki residents were allowed to vote on plans for using the budget in the Omastadi.hel.fi e-service from 1 to 31 October 2019, and the city is expected to implement the ideas with the most votes. The realization of deliberative democracy in the context of the Internet requires that the platform utilized in the e-environment is sufficiently clear and easy to use so that citizens have an equal opportunity to participate (Dahlberg 2011, p. 6). The money allocated to Helsinki budgeting is distributed by population of a district. The Helsinki city council has granted a total of EUR 4.4 million annually through PB for implementing citizens' ideas (Hel.fi 2019). The projects to be voted on are very pragmatic and are very much linked to the grassroots level of citizens’ everyday life, such as building landscapes, renovating parks and creating new meeting places for everyone. Appropriations for basic municipal activities, such as social and health services and education, are still decided by the city council.

While the amount set aside for participatory budgeting can be considered a lot in absolute terms, public investment—for example in infrastructure or public services—often proves to be expensive and, in relation to the city's total expenditure,

expenditure for participatory budgeting remains relatively low. In relation to Helsinki's total budgeted expenditure on investments (2019, 774M. €), only 0,5% of the total investment budget is earmarked for participatory budgeting. In this respect, much of Helsinki's overall budget is decided by city counsellors, which in turn undermines citizens' political involvement.

However, the low budget for PB is not the only problem that arises in Helsinki. The challenge of deliberative democracy is precisely the involvement of citizens in public debate, so that a common consensus can be reached at all. On the one hand, the problem may be citizens' indifference to "common issues". However, on the other hand, if citizens are viewed as political consumers, the city of Helsinki can be accused of poorly executed marketing that does not reach enough people to be interested in public economy. The basic idea behind PB is that it can also activate previously politically inactive citizens to participate in decision-making (Godwin 2018, p. 10-14), where Helsinki has been only moderately successful. According to the OmaStadi service, 10 435 members are involved in PB (Omastadi 2019). The distinction between active and inactive citizens is not available, but it is reasonable to assume that not all the slightly over 10,000 citizens are actively involved. Relative to the population of Helsinki (2018, 650,033), roughly 1.6% of Helsinki residents have taken advantage of their opportunity to participate in decision-making on the allocation of public funds, which, on reflection, can be considered to be a worryingly low percentage. However, participatory budgeting is still in the early stages of implementation in Helsinki, so there is reason to be optimistic about it. Budgeting as a new type of democratic innovation represents an important tool for democratizing society and involving citizens.

## 8. Conclusion

In this paper we have discussed how technology can impact citizen-governance interaction in Finland. There are many technological ways, both on the national and municipal levels, to bring governance and citizens closer to each other and to improve their relationship with each other. Nevertheless, this does not mean that the digital revolution and new technological applications have created solid solutions for the core problems of democratic deliberation or decision-making in the Finnish context.

As we first noted, it is obvious that the Internet as a whole allows a fruitful place for deliberation between citizens and governance. There are many examples, such as electronic voting, civic initiatives, crowdsourcing and participatory budgeting, which have or will have significant effects on our political and deliberative environment. These applications have, in our opinion, at least a possibility to enhance participation, communication, pluralism and accountability in governance for citizens.

It is apparent that when everyone is able to vote at home, sign civic initiatives online, or participate in crowdsourcing or municipal budgeting, deliberation and decision-making are much more transparent and also faster than they would be offline. In this situation, being an active citizen does not demand citizens' presence at any given place, at any precise time; it is flexible for both parties, all of the possibilities lie “in the same place”, and they are easier to utilise.

However, we cannot forget problems concerning anonymity, hate speech and different social bubbles which are hard to control or manage. We claim that making the Internet more friendly for citizen-governance interaction is an unfinished task that will likely continue to evolve. Technology cannot solve everything or suddenly change people or their thinking. We can further say that in Finland we are struggling with the same problems we have struggled with before. Active citizens are active online and offline, and the Internet does not automatically make passive citizens active. Second, the Internet is not a very open system, and we cannot trust technology to always work properly and solidly. There are many instances of hacking, data leakages and other similar dangers which we cannot prevent. The digital revolution is not complete.

That is why we have approached our subject in a clearly critical manner and we all claim that there are many challenges to solve before it can be even considered that technology can somehow save or enhance citizens’ ability to participate and communicate with governance. What is apparent is that technology can greatly help us in creating a better citizen-governance relationship in Finland. Time will show if participatory budgeting and civic initiatives, among other things , have an effect on decision-making and politicians’ thinking. At the very least it seems that informed citizens are likely to be more motivated to participate in issues of general interest.

## References

Aitamurto, T.,  & Chen, K. (2017). The value of crowdsourcing in public policymaking: epistemic, democratic and economic value. *The Theory and Practice of Legislation*, 5(1), 55-72. http://dx.doi.org/10.1080/20508840.2017.1282665

Aitamurto, T., & Landemore, H. (2016). Crowdsourced Deliberation: The Case of the  Law on Off-Road Traffic in Finland. *Policy & Internet* 8(2), 174-196. doi:10.1002/poi3.115.

Amoretti, F. (2006). The Digital Revolution and Europe's constitutional process. E-democracy between ideology and Institutional practices. In VII *Congreso Espanol De Ciencia Politica Y De La Administracion*. 2006, 1-18.

Bächtiger, A., Dryzek J., Mansbridge, J. & Warren, M. (2018). *The Oxford Handbook of Deliberative Democracy*. Oxford: Oxford University Press.

Brabham, D. C. (2008). Crowdsourcing as a model for problem solving: An introduction and cases. *Convergence*, 14(1), 75-90.

Büchi, R. (2011). Faktat pöytään, ihmiset pöydän ympärille. Of Ilvessalo, S. & Jaakkola, H., *Kansan valta: suora demokratia Suomen politiikan pelastuksena*, 103-118.

Dahlberg, L. (2011). Re-constructing digital democracy: An outline of four 'positions.' *New Media & Society*, 13(6), 855–872. https://doi.org/10.1177/1461444810389569

*Economist, The*. (2017).  In praise of serendipity: Social media should encourage chance encounters, not customised experiences. March 9. https://www.economist.com/books-and-arts/2017/03/09/in-praise-of-serendipity

European Commission/Kantar Public Brussels. (2018). Special Eurobarometer 477: Democracy and Elections, summary.

Floridi, L. (2014). The 4th Revolution – How the infosphere is reshaping human reality. Oxford: Oxford University Press.

Godwin, M. (2018). *Studying Participatory Budgeting: Democratic Innovation or Budgeting Tool?* https://doi.org/10.1177/0160323X18784333

Grönlund, K. & Wass, H. (2016). Poliittisen osallistumisen eriytyminen - Eduskuntavaalitutkimus 2015. Ministry of Justice, Finland.

Grossi, G. (2011). The public sphere and communication flows in the era of the Net. Of Dahlgren, The Internet, Public Spheres, and Political Communication: Dispersion and Deliberation, «Political Communication», 147-162.

Guynn, J. (Feb 13, 2019). "If you've been harassed online, you're not alone. More than half of Americans say they've experienced hate". *USA Today*.

Habermas, J. (1962). The Structural Transformation of the Public Sphere: An Inquiry into a category of Bourgeois Society.  Cambridge Polity Press.

Habermas, J., Lennox, F. & Lennox, S. (1964). *The Public sphere: An encyclopedia article.* New German Critique (3), 49-55.

Habermas, J. (1991). The Structural Transformation of the Public Sphere. An Inquiry into a Category of Bourgeois Society. Cambridge: the MIT Press.

Heywood, A. (2007). *Politics.* London: Palgrave MacMillan.

Helsingin kaupunki. (2019). Osallistuva budjetointi. https://www.hel.fi/helsinki/fi/kaupunki-ja-hallinto/osallistu-ja-vaikuta/vaikuttamiskanavat/osallisuus-ja-vuorovaikutusmalli/osallistuva-budjetointi/ 5.12.2019

Kansalaisaloite.fi. (2019). https://www.kansalaisaloite.fi/fi/ohjeet 30.10.2019

Martens, B., Aguiar, L., Gomez-Herrera, E. & Muelle-Langer, F. (2018). *The digital transformation of news media and the rise of disinformation and fake news*. JRC Technical Reports. JRC Digital Economy Working Paper 2018-02. European union.

Meyer, H. (2017). Understanding the digital revolution and what it means (blog), published Jul 06, 2017. https://www.oecd-forum.org/users/52524-henning-meyer/posts/17957-understanding-the-digital-revolution-and-what-it-means

Ministry of Justice, Finland. (2017). Working group: risks of online voting outweigh its benefits; press release 19.12.2017.

Ministry of Justice, Finland. (2017). *Online voting in Finland – feasibility study* 19.12.2017, summary.

OmaStadi. (2019). Tuo ideasi ja äänesi kuuluviin. https://omastadi.hel.fi

Shah, A. (2007). *Participatory budgeting. Washington*, D.C.: World Bank.

Setälä, M, & Schiller, T. (2012). Citizens' Initiatives in Europe: Procedures and Consequences of Agenda-setting by Citizens. United States, Palgrave Macmillan.

Spence, M. (2019). *The "Digital Revolution" of Wellbeing.* Project Syndicate, Jun 28, 2019.

Sunstein, C. (2017). #Republic. Divided Democracy in the Age of Social Media. Princeton: Princeton University Press.

Tikkala, H. (18.10.2019), Analyysi: Suostumus-kansalaisaloite tuskin etenee sellaisenaan eduskunnassa - ongelmana lapsiin kohdistuva seksuaalinen väkivalta. *Yle.* https://yle.fi/uutiset/3-11026450

Tikkala, H. (2019). Nettiäänestäminen ei ole etenemässä tällä vaalikaudella – "Vanhaan järjestelmään eivät pääse vieraat valtiot vaikuttamaan. *Yle news* 16.7.2019 https://yle.fi/uutiset/3-10880010

Väestörekisterikeskus. (2019). https://vrk.fi/en/finnish-citizens-initiative 30.10.2019

# 1.2 The Securitization of Cyberspace in the Finnish Cyber Security Strategy 2019

Riikka Pasanen, Juho Mölsä, Noora Magd, Eetu Kukila, Sami Husa
Faculty of Social Sciences, University of Helsinki

## Abstract

The securitization of cyberspace is the process through which actors declare elements of cyberspace as fundamentally security issues, often coming into conflict with fundamental values of deliberative democracy, such as freedom of expression and the right to privacy. This paper will examine the securitization of cyberspace in the Finnish context, using Thierry Balzacq's framework (2011) of securitization, in the context of the newly released Finnish Cyber Security Strategy (FCSS) of 2019, and address the discourse on threats surrounding the creation of the strategy, as well as contextualizing it in the EU's collective securitization while paying particular attention to how actors with limited technological autonomy, such as Finland and Estonia, formulate their strategies. The paper will make policy recommendations for finding the right balance between security and consideration for privacy and human rights that allow deliberative democracy to flourish.

*Keywords:* Securitization, cybersecurity, Finland, Balzacq, Copenhagen School, Finnish Cyber Security Strategy, FCSS, Estonia

This research paper aims to study the securitization of cyberspace in the context of Finnish cyber security. Our paper seeks to explore how cyberspace has been securitized, what threats have emerged through the process, and what policy options have been or should be implemented to combat the threat in a Finnish context. Cyber security is a current phenomenon in Finland and, as we show below, has been the subject of concrete political actions and public strategies that aim to address the challenges related to the domain. Our case study approach aims to provide new comparative insights on the subject and highlight the topicality of the issue. Furthermore, a case study approach allows us to provide more detailed information on different timelines, actors, and possible policy actions to combat issues arising through the unfriendly use of cyberspace and account for the different interests of the writers.

Section One begins with situating the research framework within securitization theory and cyber domain literature. The theoretical framework consists of Balzacq's securitization theory and the sociological view of securitization which allow us to identify the processes through which threats become a politicized issue while a brief discussion of the concepts of cyberspace and cyber domain help to understand the strategic processes of securitization in the given domain.

Section two analyses the Finnish Cyber Security Strategy (FCSS) and its socio-historical context while section three applies Balzacq's case study framework to the FCSS. Section two's timeline encompasses past, present and future, which provides insight into the development of cyber security in the Finnish context. The focus on agents and practices in section three serves to highlight the relationship between co-optation and regulative practices in Finnish strategy.

Section four contextualizes the development of the securitization of the cyber security strategies of Finland and Estonia in the 2007 Estonia cyberattacks. Section Five ties the use of cyberspace to hybrid interference, an element of threat to the tenets of deliberative democracy, and briefly discusses possible responses and their limitations. The aim is to recognize policy recommendations varying from resilience to democratic deterrence. Finally, the conclusion summarizes the research and lays the groundwork for further and more in-depth investigation of this multi-faceted issue.

## Situating the framework in securitization and cyber domain literature — Eetu Kukila

This research paper takes the theoretical framework laid out in Thierry Balzacq's theory on securitization as its starting point. This section outlines the theoretical basis for understanding the cyber domain to facilitate an understanding of the role of cyber security and the merits of its study in general. We begin with a brief overview of securitization theory before focusing more closely on the sociological view of securitization. Lastly, the section reviews scholarship on the cyber domain with a focus on how it is understood in the security context.

## Securitization theory and its value

The securitization of cyberspace is a process by which actors declare elements of cyberspace as a fundamental security issue. Balzacq offers a framework for securitization, i.e. something becoming a security issue through discursive politics. Moreover, securitization is an especially intersubjective process, which transforms an abstract threat into a concrete being. Furthermore, security problems can emerge from different practices, including sociological and philosophical, but their original intention was not to create a security problem. (Balzacq 2011, 1-3).

Securitization processes are initiated through speech act expressions, which are more than statements. Rather these representations are performatives; they can take actual actions and result in concrete effects on the environment. Put it another way, they are not "constatives that simply report state of affairs and are thus subject to truth and falsity tests" (Balzacq 2011, 1, 15). The Copenhagen School has laid the foundation for speech act on which Balzacq has continued to develop it, expanding the scope of the study to policy tools. In speech acts, there are always two actors: the speaker and the audience. In order to be successful in securitization, it is required that the audience accepts the reality of the threat (Peoples & Vaughan-Williams 2010, 78).

Securitization theory contains three assumptions: the centrality of audience, the dependency between context and agency, and forming practices (Balzacq 2011, 3). When trying to widen the security agenda there lies a political danger in tacking the word "security" on to a wide range of issues. Therefore, the main concern is to define what is part of security, and what is not in the context of a broadened understanding. (Peoples & Vaughan-Williams 2010, 76). Similarly, security has been expanded to include a variety of issues in the cyber domain. Crucially, when the issue comes to be treated as a security matter, it is justifiable to use exceptional political measures to deal with it. Thereby an issue becomes securitized when it morphs from non-politicized through politicized process to become securitized (Peoples & Vaughan-Williams 2010, 77).

Securitization theory offers us a research framework for how to identify and understand threats. Public attention on the issue and the requirement for legal and political action are the key criteria when operating on security matters. As Balzacq aptly puts it: "the *shared salience* of an issue marked by the *imperative of acting now*, constitute necessary and sufficient conditions for securitization" (Balzacq 2011, 32). Also, as in our research, securitization theory takes the form of case studies constituting the primary research strategy. Hence, empirical inquiry investigates real-life context wherein the audience has a direct connection with the security issue and an ability to take actions to counter the threats (Balzacq 2011, 32-34).

## A sociological view of securitization

When doing a case study through the framework of securitization there must be scrutiny on the practices of how policies are created, and which kind of tools have been deployed. Additionally, context should not be ignored, as it poses a threat to the historical and social environment. (Balzacq 2011, 35-36). Balzacq thus offers two different approaches to securitization: sociological and philosophical. In our research, the sociological view is more useful and thus will be presented more broadly.

Accordingly, the sociological model relies primarily on practice and context. It views securitization as a strategic process affected by context and the disposition of the audience. The strategic action of a speech act works through interaction while trying to make security issues more open to universal communication. In other words, persuasion transfers the security issue from the non-political area to the political through a securitizing process. The sociological model stresses the actors' habitus, which can be understood as their engrained historical characteristics that inform others about their behaviour. Habitus may vary depending on context and therefore on changing historical situations. In addition, the audience and actors have an equal role in constituting securitization. (Balzacq 2011, 1-2).

An overview and basis for the framework of securitization theory has been given above. Securitization theory offers an analysis of the cyber domain and its relation to security problems. Cyberspace contains a variety of different security issues that may provide new insights when combined with securitization theory. Next, I will focus on theorizing cyberspace and the security issues related to it.

## Cyberspace and security

The reviewed cyber domain literature is oriented to the study of cyber security in a defensive context. The literature often surveys different policies of a variety of sovereign states on how to respond to cyberattacks. For example, Finland's National Defence University provides researched information on the security aspects and sovereignty of a cyber domain in their research called The Fog of Cyber Defence (2013). Different studies are concerned with what steps states must take to protect their cyber domain.

To study the cyber domain, one first needs to understand what cyberspace is. When theorizing cyberspace, it can be called "cyber-space-time", because it allows information to travel instantly at great distances causing a compression of time and space. In other words, cyber-space-time manipulates time with cyber technology. According to complexity theory, cyberspace is complex since individual actors spontaneously interact with themselves, making outcomes unpredictable. Cyberspace is basically accessible to everyone, because no actor has a natural advantage in the cyber domain, and it is separate from other domains. However, the performative action in cyberspace can amplify different capabilities in other domains (Russell 2014, 12-13). Thus, cyberspace is an important tool for other domain functions as well.

Like any other domain, cyberspace also needs security, and securing sophisticated technology by relying on similarly sophisticated technology is regarded as a challenge. Thus, ensuring the adequacy of cyber security technology has become problematic. There is a risk, however, of reducing the human factor excessively. When considering cyber security, it requires human beings and their ability to use cyber domain for their own benefit. (Rantapelkonen & Kantola 2013, 31-32). So real human actors must be kept in mind as the end-users of cyberspace.

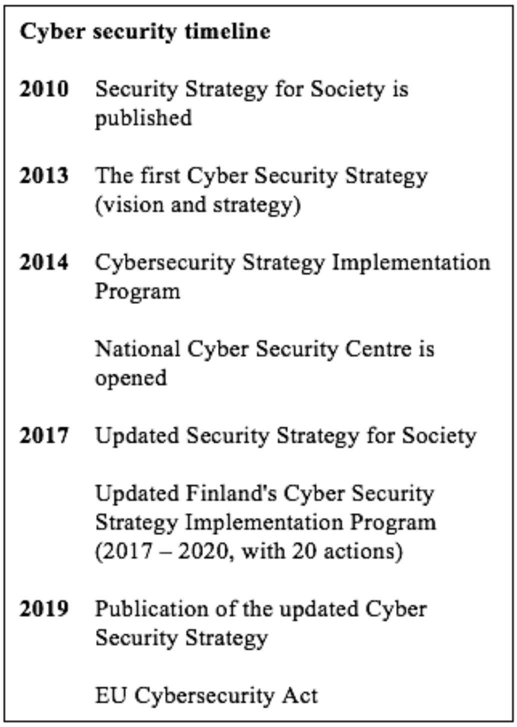
Cyberspace is not a neutral environment and it is divided among many different actors who can be strong or weak in relation to one another. In addition, cyberspace contains a complex web of interdependent actors that reflect world politics (Rantapelkonen & Kantola 2013, 32). Therefore, cyberspace is a problematic area when considering legislation. Existing laws and norms are not yet developed as workable and credible enough to deal, for example, with warfare related issues in cyberspace (Russell 2014, 13). Scholars have agreed that without specific rules of international law, the cyber domain located in the jurisdiction of a state is subjected to administration by the respective state (Tuukkanen 2013, 42). This can lead to serious security dilemmas when traditional deterrence strategies from other domains are adapted to the cyber domain (Russell 2014, 14).

## The past, present and future of the Finnish Cyber Security Strategy (FCSS) — Riikka Pasanen

This section will examine the development of policymaking on cyber security in Finland. Finland is an advanced information society whose activities in both the private and public spheres depend on various electronic networks and the services they provide. This vulnerability to external threats rises from the dependant relationship with information technology (IT). As defined in Section 1, cyberspace is a contested territory and an arena for political maneuvers, including power plays and acts of war. Potential threats posed by cyberattacks to modern information society include computer hardware and systems failure and collapse of information infrastructure, which have a negative impact on public services, business and administration and therefore on the basic functioning of society (Lehto et al. 2017). Many of the interconnected societies of today are increasing their cooperation as these vulnerabilities are shared by most, if not all information societies. This development of increasing internationality is visible in the evolution of the Finnish Cyber Security Strategy even if its history spans only little more than a decade, and is foreseen to continue. In this section, the roots and context of the FCSS are explained, taking into account the socio-historical context, and a brief summary of recent developments, as well as a look to the future, are provided.

### Emergence and historical context of the Finnish Cyber Security Strategy

When it comes to nations grappling with the need to protect their borders, the years 2007 and 2010 can be seen as major turning points with the Estonian cyberattacks and the discovery of the Stuxnet worm, respectively. In the FCSS, only the latter is mentioned, nevertheless, it would be imperceptive to suggest that the Estonian case and the following process of 'hypersecurization' (see section 4) would not have been extremely closely followed in neighboring Finland. The discovery of Stuxnet, an incredibly effective malware program, is referred to as "*the beginning of a new era in cyber security*" in the first FCSS (the Security Committee 2013, 18). Stuxnet is infamous for enabling an attack on Iran's nuclear facilities and physically damaging centrifuges for uranium enrichment. The code was developed secretly, but experts agree that developing it must have required incredible expertise and likely more than one nation pooling their resources. The damage caused by this single, powerful cyber worm delayed Iran's nuclear development for years and showed the international audience that cyber tools are able to cause physical damage to electronic devices and systems.



**Cyber security timeline**

| | |
|---|---|
| **2010** | Security Strategy for Society is published |
| **2013** | The first Cyber Security Strategy (vision and strategy) |
| **2014** | Cybersecurity Strategy Implementation Program |
| | National Cyber Security Centre is opened |
| **2017** | Updated Security Strategy for Society |
| | Updated Finland's Cyber Security Strategy Implementation Program (2017 – 2020, with 20 actions) |
| **2019** | Publication of the updated Cyber Security Strategy |
| | EU Cybersecurity Act |

The beginnings of the FCSS could be argued to be found in 1995 when the government of Finland first published a national guideline, "the Finnish Information Society Strategy". However, the ramifications of a cyber influenced reality were likely not foreseen very clearly at this early stage of the Internet. For the purpose of this project, the 2003 Government Resolution on the Strategy for the Protection of Critical Activities in Society (YETT) is a more reliable starting point for tracking the predecessor strategies of the FCSS. YETT, which focused on securing vital functions of society, was last updated in 2006. During the 2010 update, the name was changed to better reflect the contents of the document. By the time of publication of the Security Strategy for Society in 2010, cyber security had been noted to be an integral part of national security (Ministry of Defence 2010). Finland had already been the target of cyber operations, with a focus on cyber-activism, crime and espionage (the Security Committee 2013), which has continued since, and is now an inherent feature of an information society. In the Security Strategy for Society, it was noted that Finland is lacking in cyber security, especially in the area of coordination of response (Ministry of Defence 2010). Less than a year later, in March 2011, the decision was taken to begin the preparation of the FCSS as a direct outcome of the Security Strategy for Society (the Security Committee 2019).

The first Finnish Cyber Security Strategy is an ambitious document, outlining its vision for the future as "*By 2016, Finland will be a global forerunner in cyber threat preparedness and in managing the disturbances caused by these threats"* (2013, p.3). The publication of the first Finnish Cyber Security Strategy provided a road map for developing Finland's competencies further and further securing the "national" cyberspace, leading up to the launch of the National Cyber Security Centre, the first new program implemented in the scope of the FCSS and the subsequent Cyber Security Strategy Implementation Program that formatted the goals outlined in the strategy into a set of concrete steps and policy actions. The key tools for managing cyber security listed in the first Finnish Cyber Security Strategy (2013) are *prevention*, *detection* and *capacitation.*

A review was commissioned by the Government to assess the impact of the FCSS in anticipation of the upcoming revision of the FCSS. The researchers found advances in the field of cyber security, but also that significant measures were needed for Finland to "establish a global forerunner position in cyber security" as that had originally been set as a target by 2016. However, the review did find that Finland is consistently in the top of indexes comparing cyber security preparedness and also noted that each of the other nations compared in the study, including Estonia, the Netherlands, and Singapore were likewise positioning themselves for positions of global or regional forerunner in the field of cyber security (Lehto et al. 2017, 62).

### A renewed focus for a new decade: streamlining leadership and increasing international cooperation

The program outlined in the Cyber Security Strategy Implementation Program contains 74 actions such as policy changes and new internal steering mechanisms, providing a framework for the development of national cyber security with concrete measures. Perhaps the most known implemented action from the program was the launch of the National Cyber Security Centre in 2014.

A key part of concrete FCSS implementation has been "deciding who gets to decide". One of the observations of the 2010 Security Strategy for Society was the division of jurisdiction between government departments and other entities, which was counterproductive to the perceived need of clear, expert consultancy and even co-ordination in future crisis situations. In the midst of governmental pressures to streamline, reformation of jurisdictional borders between agencies has also been employed. This was likely partly in response to the budgetary pressures to streamline government operations but also due to the unpredictable development of cyber threats such as hybrid interference (see Section 5). A result of this process is the Security Committee, an interdepartmental coordinating body within the Ministry of Defence, tasked with assisting all the Ministries. (the Security Committee, n.d.) Compared to its formal predecessor, the Committee for Security and Defence, the new Security Committee is broader in its scope and the membership of the committee has been consequently amended to reflect the expanded purpose of the Committee.

The brand new FCSS 2019 sums up its three core areas as leadership, capacitation and international cooperation. Whereas the FCSS 2013 outlined an incredibly ambitious vision of Finland as a global forerunner in cyber threat preparedness and in managing the disturbances caused by these threats, the revised edition places a greater emphasis on the international cooperation aspect, also owing to the 2019 EU regulation, the Cybersecurity Act and the consequent pooling of resources in the form of the European Cybersecurity Industrial, Technology and Research Centre and setting up a network of National Coordination Centres (European Union 2019). On the national level, the topic of cyber security in the government will gain more visibility with the upcoming new appointment of a Cybersecurity Director. This position was decided to be established within the Ministry of Transport and Communications to coordinate the national development of cyber security (the Security Committee 2019).

## Co-optation of economic interest for securitizing purposes — Agents and practices in the Finnish Cyber Security Strategy 2019 — Juho Mölsä

This section will apply Balzacq's case study framework (2011, 32–38) to the Finnish Cyber Security Strategy (Government of Finland 2019). As the context has been presented in the second section, I direct my attention to agents and practices. Since our case study is a strategy including policies and practices, I am using what can be called a sociological (Balzacq 2011, 22) or a practice-oriented approach (Balzacq, Léonard, & Ruzicka 2016, 504–507).

### Co-optation of economic interest and national competitiveness

In line with Avant and Haufler's (2018) findings, the Finnish Cyber Security strategy challenges classical state-centered views of security, at least partially. Private actors are seen not only as referent objectives but also functional actors (Balzacq 2011, 7), as those who protect against the threats (Government of Finland 2019, 7, 9). The securitization of cyberspace is even framed as advantageous to "national competitiveness" (Government of Finland 2019, 8). This linkage of economic and political interests raises questions on whether we should classify some private actors as securitizing actors, which I will return to below. It also offers one hypothetical motivation for the securitization of Finnish Cyberspace: national competitiveness.

Although the securitization literature has reserved a central role for states or public actors (Avant & Haufler 2018, 3) Hansen & Nissenbaum (2009, 1161–1162) find that in the cyber sector, private actors have been vested with responsibility due to the networked nature of the cyber domain. Hansen and Nisselbaum (2009, 1162) argue that discourses in cyber-related securitization can be analysed as constellations of multiple referent objectives that can include both private and public, economic, and political referent objectives. In our analysis, we make similar observations using our practice-oriented approach and recognize different elements as part of a common *dispositif* (Balzacq et al. 2016, 505).

In the Finnish Cyber Security Strategy there is a "co-responsibility" (Hansen & Nissenbaum 2009, 1162) on cyber security between private and public actors where some of the responsibilities are given to private actors (Government of Finland 2019, 7–9). The private responsibilities include, for instance, maintaining vital infrastructure and offering security services for businesses (Government of Finland 2019, 7–9). This can be interpreted as a co-optation tactic in light of the regulative instruments and capacity tools that I describe in more detail below. In addition, the above-mentioned linkage with national competitiveness can also be interpreted as a tool for co-opting economic interest groups for securitization purposes. The use of co-optation tactics supports a key finding in practice-oriented securitization studies, the intersubjectivity of securitization (Balzacq et al. 2016, 469, 499), as it makes it difficult to separate the audience from the agent.

### Epistemic community, capacity building, and regulative powers

Despite some of the functional, securing, actors being private, it is still true that state authorities are identified as the key security actors (Government of Finland 2019, 5–6). Avant and Haufler (2018, 9–10) argue that states have reserved the role of commissioning and legitimizing private actors in the security field. Thus, they can be classified as the securitizing agent. In the case of the Finnish Cyber Security Strategy, this can be seen most clearly in the obvious fact that the strategy is a government-issued document (Government of Finland 2019, 1). Furthermore, the division of labor is implicitly noted as the partnerships and modes of action are to be bound by regulation (Government of Finland 2019, 4).

Critical securitization literature has identified a “military-industrial-bureaucratic-scientific complex” (Eriksson 2001, 213) or in Haas’ words an “epistemic community” (Eriksson 2001, 215) as securitization agents. The Finnish context appears similar, as the Finnish Cyber security strategy has been produced by a constellation of state bureaucrats, military personnel, and experts from the field (Security Committee 2019). This complex appears as the key agent in the implementation as the practices in the strategy propose resources towards research and development (Government of Finland 2019, 9), government coordination exercises (Government of Finland 2019, 5) and security authorities (Government of Finland 2019, 6). The choice of a practice-oriented securitization approach allows us to take the production of the report itself into account as an example of practice.

Eriksson (2001, 219) in his study of cyberspace securitization in Sweden, notes that the Swedish case framed the threat as “cyberwar” (Eriksson 2001, 219). Thus, the role of the private sector was “restricted” although private actors such as business associations were involved in cyber security discussions (Eriksson 2001, 219). The Finnish Cyber Security Strategy has a whole-of-society resilience and business orientation approach (Government of Finland 2019, 4, 9) and thus, private interests are naturally more involved. For instance, the group that participated in the preparation of the Finnish Cyber Security Strategy by submitting official statements included business associations (e.g. the Confederation of the Finnish Industries) and some private cyber security companies (e.g. F-Secure) (“Lausuntopalvelu.fi” 2019). As noted above, businesses are also heavily involved in the practices proposed in the strategy. Thus, representatives of private economic interests are active members of the epistemic community in the Finnish context.

In light of the constellation of the epistemic community, it is worth noting that there are objectives concerning public actors that are "regulatory instruments" (Balzacq 2011, 17) but the objectives concerning private actors are mainly "capacity tools" (Balzacq 2011, 17). Regulative instruments include having up-to-date capabilities and toolboxes for authorities which includes proposed new legislation empowering authorities (Government of Finland 2019, 7). Capacity building, on the other hand, includes tools such as conducting knowledge sharing exercises (Government of Finland 2019, 9) and financial support, including private, training, research and technological development (Government of Finland 2019, 9).

The last point on the capacity-based approach returns us back to the co-optation tactics. In sum, a slightly paradoxical approach appears in the effort to create a culture of cyber security together with the whole-of-society, but at the same time empowering authorities with new resources and capabilities.

### Conclusions

To conclude, I would like to critically assess the chosen securitization approach. In this section I have used securitization in a broad sense, meaning that the utterance of security or identifying disastrous threats in certain fields is classified as securitization (Balzacq et al. 2016, 503). This clearly applies to the Finnish Cyber Security Strategy. The agents and practices analysed in this section would also support, at least partially, another sufficient condition that cyberspace is seen as a field of security authorities, such as military and security police (Eriksson 2001, 219). Of course, it is a legitimate question to ask to what extent economic agents and rhetoric affect the interpretation of the correct sector of the strategy. However, there is at least one, more narrow condition for securitization regarding practices that the Finnish Cyber Security Strategy does not fully support. The condition is that the practices should be extraordinary and not fully comply with normal democratic procedures (Eriksson 2001, 220). This is not the case with the strategy since all the proposed new powers for authorities are meant to be formally adopted following normal framework development and legislative procedures (Government of Finland 2019, 4).

## Comparison: How actors with limited technological autonomy operate; the Finnish and Estonian Cyber Security Strategies — Sami Husa

### Introduction

The denial-of-service cyber-attacks against Estonian government institutions, telecommunications, banking and infrastructure in April and May of 2007 are described as "the first real war in cyberspace" by Hansen and Nissenbaum (2009). The upshot of the attacks was the drafting of Estonia's first Cyber Security Strategy (ECSS), now in its third iteration. The attacks similarly jolted Finland into action as discussed previously, leading to an urgency in drafting policy and setting and providing a plausible threat scenario for it.

When Estonia launched the ECSS in 2008, it was one of the pioneers of the domain, with only three other countries having produced a cyber security strategy at the time (Pernik & Tuohy 2013). The development of Finland's first Cyber Security Strategy (2013) reflects the same strategic priorities as those of Estonia; how a state with limited technological autonomy can respond to cyberattacks, the development of a horizontal model of resilience, and the fusion of private, technical and national interests. Securitization theory can help conceptualize the politics surrounding the development of cyber security policy and how interactions between the securitizing actors and the audience legitimize these policies.

### Hansen and Nissenbaum's theoretical framework of cyber security

Hansen and Nissenbaum's seminal 2009 paper on cyber security proposes three security modalities in the practice of cyber-security; *hypersecuritization*, *everyday security practices* and *technifications* (Hansen & Nissenbaum 2009). Hypersecuritization is presented as the process of distorting and exaggerating the enormity of security threats; everyday security practices as the creation of individual and business 'compliance in protecting network security' and familiarization with threats, and technification as the creation of a privileged space and authority for experts with technical knowledge, outside of the realm of politics (Hansen & Nissenbaum 2009). In their reading, hypersecuritizations always "mobilize the spectre of the future" while also using "the past as a legitimating reference" (2009). This is demonstrated in the third Estonian Cyber Security Strategy 2019 - 2022 (ECSS), which invokes the 2007 attacks repeatedly from the introduction onwards. The Finnish Cyber Security Strategy 2019 (FCSS) is also broadly security and threat-focused; indeed, the FCSS 2019 is entirely devoid of references to protecting privacy or the cyber realm from militarization, and only very briefly mentions the opportunities of digital technology to society (FCSS 2019).

### The 2007 Estonia attacks and the creation of cyber war

As outlined by Robert Kaiser (2015), the 2007 Estonia cyberattack, precipitated by the removal of the Bronze Soldier in Tallinn, catapulted Estonia from a position on the margins to the epicenter of Western cyber security discourse. NATO's Cyber Defence Center of Excellence was established in the city the very next year, and the Estonians continue to provide an outsized influence on EU policy on cyber security (Pernik & Tuohy 2013). Kaiser argues that the ensuing militarization of Estonian cyber security discourse led to an enclosed circle of knowledge among officials, and a "depoliticized discourse of unquestionable truth", or *techinification* in Hansen and Nissenbaum's framework (Kaiser 2015).

As Arquilla points out, the ensuing hypersecuritization of Estonian cyberspace was excessive compared to the reality of the threat (2012). While Estonian Prime Minister Andrus Ansip likened the attacks to a blockade, and publicly discussed invoking NATO's mutual defense clause, Arquilla argues that the attacks were ultimately little more than a nuisance; the attack was non-violent, it carried no explicit demands for Estonia to change its political behavior, and no state actor or political entity took credit for them (Arquilla 2012).

The 2007 cyberattacks and ensuing securitization can be understood as a speech act exposing the ways in which "Russianness" performatively materialized as the outside of "Estonianness" (Kaiser 2015). In this reading, the denial of service attacks was presented as an act of war threatening Estonian sovereignty by Russia, aided by a fifth column of Estonian-Russians. This made it possible to project the "spectre of the past" on current events, by imposing the familiar Cold War narrative into the cyber domain. This threat also afforded an opportunity; as Kaiser argues, the frame of cyber war allowed the state to recast the troubled relationship between Estonian nationalists and the state's Russian minority, to present Russia as a renewed threat to US and EU allies, and to establish its own role as a "transactor" in cyber security (Kaiser 2015).

### Determining audience assent

The securitization of an issue does not depend on objective events but rather stems from the interactions between the securitizing actor (the agent who presents the threat) and the audience (Balzacq et al. 2015). For a concept so central to security theory (ST), determining what constitutes an audience and how its assent is determined has been one of the least developed concepts in ST (Balzacq et al. 2015). The Finnish and Estonian Cyber Security Strategies (CSSS) articulate a threat and a response, and are therefore securitized, but how is audience acceptance quantified? Roe (2008) points out that there is not a single audience, but a multitude. Using the example of the decision of the UK government to join the Iraq war, Roe argues that while the formal agreement of one audience (Parliament) was received, wider public opinion was against the war. Similarly, the creation of the Finnish and Estonian CSSS demonstrates the acceptance of securitization on a political-bureaucratic level, as discussed previously, but does not as such consider audience assent, outside of stressing social resilience as a defense against threats.

According to the Estonian Ministry Defence annual public opinion survey, 75% of Estonians rate cyberattacks as the most likely threat to their country, marking it as the threat Estonians are most concerned with (Estonian Ministry of Defence 2019). This is consistent with the results of public opinion surveys since 2006, with cyberattacks coming in either as the first or second highest rated area of concern (Veber & Bloom 2016). The Finnish Ministry of Defence's most recent public polling quantifies cyber-threats as the 5th largest area of concern for Finns, with 79% perceiving cyber security attacks as very concerning or concerning, a steady growth from 2009 (Finnish Ministry of Defence 2018). While the opinion polls serve imperfect measures of audience assent, they indicate a successful securitization of cyberspace, particularly in the case of Estonia.

### Conclusion

As the Finnish and Estonian cases of securitization demonstrate, securitization is essentially a political process, and politics cannot be removed from theory by deriving it from objective threats (Wæver 2011). At the same time, the theory has an implicit preference for desecuritization, by fostering critical analysis of the potential of securitization in itself to distort or generate representations of threats (Wæver 2011). Concurrently, the reality of the risks outlined in the strategies can be underplayed when viewed through securitization theory. Heikki Patomäki (2015) offers a criticism of using speech act theory to explain security threats, and the ambivalence the theory has towards being able to predict future risks meaningfully. While the future is unknown, he argues that there is a rational method for assessing the probability of possible futures. Therefore, the dramatization of threats to democracy, infrastructure and systems integrity presented in the strategies may be considered sensible (Patomäki 2015). When a particular future threat is not only objectively real but also probable, the best policy response to the danger is a political question of great importance. The next section will consider some of these scenarios, and offer policy recommendations.

## Hybrid interference through cyberspace — Threat and action — Noora Magd

Following the securitization of cyberspace mapped out in the previous sections this part of the research paper focuses on the role of cyberspace as a vector through which to conduct hybrid interference. Actions taken by governments against hybrid interference, whose effect is dependent on its subtle nature, and especially its application through cyberspace run the risk of over-securitizing society and acting against democratic states' aims of upholding liberal values. At the same time not responding to attempts aiming to erode the basis of democratic societies run the risk of allowing deliberative democracies to crumble.

Responses to this dilemma have been offered in terms of resilience and a whole-of-society approach tied to the notion of democratic deterrence (Wigell 2019). More pessimistic approaches, however, regard the West and Finland as having already lost the game in regard to an open cyberspace due to Russia's process of constructing a closed Internet (known as RuNet) and the inherent asymmetry it presents (Kukkola et al. 2017).

### Hybrid interference – the threat

Defined as the application of non-military tools to clandestinely influence a state and its society, hybrid interference is a tactic tied to a strategy of wedging, where the synchronized use of different methods of disruption is applied in an attempt to bring about deep divisions within the target state's society (Mikkola et al. 2018). This strategy of wedging aims to disrupt political processes within countries, undermine trust in public institutions and sabotage the social cohesion of liberal democracies through division (Wigell 2019). It can be argued that the same tenets that uphold liberal democracies (limited public state power, pluralism, freedom of press and information, as well as an open financial market with a limited ability to monitor civil society) are its inherent weaknesses allowing for hybrid interference to occur and making it harder to detect (Mikkola et al. 2018).

It is essentially this pluralism and freedom of press inherent to liberal democracies and deliberative democracy that is threatened by hybrid interference tactics applied through cyberspace. The proliferation of social media and the widespread use of the Internet has widened the area of public deliberation and enabled previously marginalized positions to come forward. This has enhanced the participation in public deliberation and activism in the form of citizens' initiative, for example. At the same time, the expanse of social media has also been exploited by state and non-state actors, such as Russia, to skew public discussion through the use of trolls and bots, using algorithms to influence opinions and financing right-wing and other extreme groups to strengthen already existing divisions within society. (Mikkola et al. 2018).

The public sphere, so critical to deliberative democracy according to Jürgen Habermas (1991) and its champions, e.g. journalists, have increasingly become targets of hybrid interference. For example, Jessika Aro from the Finnish Public Broadcasting Company YLE has become a target due to her investigation of Russian trolls and bots. To the date, a Finnish docent and the editor in chief of MV-magazine, both known as pushing a Pro-Russian agenda, have been tried in cases related to intimidating Aro (BBC 2018). MV-magazine has also been a large contributor of anti-immigration views, bringing about a polarization of politics and strengthening divisions within Finnish society while receiving funding from Russian sources. Most recently Russian linked sources have been connected to actions aiming to influence the European elections (European Commission 2019). As in the case of Aro, Russian approved, conducted or facilitated actions to silence critics through cyberspace can take many forms, including, but not limited to cyber espionage, the use of trolls, fake news and smear campaigns as well as threats (Aro 2019).

The erosion of a civic culture that includes deliberation due to the polarisation and restriction of freedom of press through intimidation and flooding social media with fake news is a risk to democratic governance that rests on such conditions to prevail and reach consensus (Mikkola et al. 2018). Thus, the use of cyberspace as a tool of hybrid interference presents a challenge to open liberal democratic societies.

### Democratic deterrence and resilience – the response

The main challenge lies in how to respond to these actions conducted in cyberspace. Western states can be said to have been forced into acting reactively when engaging with these threats thus far. Newer policies such as the EU joint communication on fighting disinformation, the new Finnish cyber security strategy and the new Finnish intelligence laws (European Commission 2019; the Security Committee 2019) aim to combat these challenges more proactively.

A framework of democratic deterrence proposed by Wigell includes notions that any response action must rest on the liberal values associated with democratic states. Democratic deterrence thus includes a whole-of-society, soft power approach to cyber and hybrid threats that includes non-military asymmetrical means of response that serve a restricted security aim (Wigell 2019). Democratic deterrence is based on the notions of the attractiveness of democratic values, the commitment of democratic states to liberal values and the instrumentalization of non-state actors within society and instruments such as transparency, rule of law and citizen activism to reveal acts of hybrid interference and deter illegal actions (Wigell 2019).

Looking at the European Union's inability to respond to the crisis in Syria and Ukraine casts doubt on the inherent attractiveness of democratic values, the use of soft power and thus "democratic compellence" (Wigell 2019). At the same time, democratic deterrence's incorporation of civil society into a strategic action plan and the capability to respond in a different domain, from where the attacks occur do offer avenues of exploring possibilities to combat threats and attacks emerging through cyberspace.

A key element raised in strategies, such as the whole-of-society approach, the FCSS (2019) and the European Union's Global Strategy (European Commission 2016) is promoting resilience within critical infrastructure, services, states and the union. Resilience-based strategies can be divided into three categories: 1) defensive/gaming to keep up the status quo, 2) marginalizing/cordoning off the threat, and 3) open/rejuvenating strategy (Mikkola et al. 2018). The FCSS thus far seems to incorporate a combination of the first and second strategies while tying into a larger notion of a whole-of-society approach to security championed by Finnish actors as a whole (see for example the Finnish Security Strategy for Society 2017). What strategy is chosen decides the framework of action and how the threats represented in cyberspace or through hybrid interference at large are conceptualized. In the words of Romeo Dallaire, the question is whether "we want to survive or thrive?".

Kukkola et al. (2017) explore the possibilities of a closed Russian Internet and its effect on Western states. The result is an inherent asymmetry in cyberspace that is difficult to overcome. Closing our own networks to respond to the challenge of closed Internets would result in a win for autocratic regimes due to the end of an open public Internet embodying western liberal ideals. Keeping our networks open, on the other hand, leaves us vulnerable to continued cyberattacks and hybrid interference, resulting in a possible victory for autocratic regimes. (Kukkola et at 2017).

Previous studies on Russian interference in the US and French elections seem to point towards a decreasing effectiveness of tools of hybrid interference in successive applications and the role of media literacy in the population. This would suggest that policies incorporating resilience enhancing measures and the whole-of-society approach may be the best approach to combating cyber operations and acts of hybrid interference through cyberspace without closing down an open Internet and thus compromising on democratic values.

# Conclusion

This paper considered the securitization of cyberspace in the Finnish context from a wide variety of perspectives and drew conclusions from them. In Section 1, we provided an overview of Balzacq's securitization theory, which posits securitization primarily as a speech act that sees securitization as essentially the process of posing an exceptional threat and persuading an audience of the reality of this threat. The section also considers how to define cyberspace and the unique challenges of securing a complex web of interdependent actors lacking in international norms. In Section 2, the development of the Finnish Cyber Security Strategy is presented in its historical context, outlining the evolution of Finnish policy towards cyber security from 1995 up to the FCSS 2019. The section outlines how Finland's ambitious policy documents have belied the lack of budgetary and political support actually received. Section 3 considers the constellation of public and private actors in the formulation of the FCSS while arguing that agents and practices in the Finnish context support a broad definition of securitization not fully aligned with some of the narrower theoretical conditions. Section 4 traces the genesis of both the Finnish and Estonian cyber security strategies in the founding myth of the first "cyberwar" of 2007. The section critically examines the speech act that led to securitization, and how it is used to justify policy hypersecuritization. Section 5 offers another viewpoint, moving from the conceptualization of securitization as not only a reaction to an outside threat but as the goal of hostile actors. The section argues that hybrid interference is best countered with a whole-of-society approach, balancing the need for security with respect for liberal democratic values.

Examining the Finnish Cyber Security Strategy through the framework of securitization theory calls for further development of the theory beyond a focus on exceptional measures. As we have argued, the securitization of cyberspace has led to a logic of security that relies on resilience and a fusion of technical and national security. In this context employing exceptional measures would be a sign of strategic failure. We believe this paper has identified several promising leads for research;

- Can there be a normative theory for cyberspace securitization that sets out the criteria for when a threat is genuine?
- To what extent can the decreasing efficacy of hybrid warfare actions be attributed to cyber security strategy?
- How does the establishment of a national version of the Internet by authoritarian regimes affect it as a global platform for exchanging ideas and a global deliberative space?

## References

Advisory Board for Defence Information (2018). ‘Finns’ Opinions on Foreign and Security Police, National Defence and Security’.’’

Aro, J. (2019). *Putinin trollit*. Helsinki: WSOY/ Johnny Kniga

Arquilla, J. (2012). ‘Think Again: Cyberwar’, *Foreign Affairs*, 27th February: http://foreignpolicy.com/2012/02/27/think-again-cyberwar/

Avant, D., & Haufler, V. (2018). Public–Private Interactions and Practices of Security. *The Oxford Handbook of International Security*. https://doi.org/10.1093/oxfordhb/9780198777854.013.23

Balzacq, T. (Ed.). (2011). *Securitization theory: How security problems emerge and dissolve*. Milton Park, Abingdon, Oxon; New York: Routledge.

Balzacq, T., Léonard, S., & Ruzicka, J. (2016). ‘Securitization’ revisited: Theory and cases. *International Relations*, *30*(4), 494–531. https://doi.org/10.1177/0047117815596590

Balzacq, T., Guzzini, S., Williams, M. C., Wæver, O., & Patomäki, H. (2015). ‘What kind of theory – if any – is securitization?’ *International Relations,* 97-102

BBC. (2018). *Jessikka Aro: Finn jailed over pro-Russia campaign against journalist*. Accessed 24.10.2019 https://www.bbc.com/news/world-europe-45902496

Eriksson, J. (2001). Cyberplagues, IT, and Security: Threat Politics in the Information Age. *Journal of Contingencies and Crisis Management*, *9*(4), 200–210. https://doi.org/10.1111/1468-5973.00171

Estonia Ministry of Communications (2019). Cybersecurity Strategy 2019 – 2022.

Estonia Ministry of Defence (2019). ‘Public Opinion and National Defence’.

European Commission, High Representative of the Union for Foreign Affairs and Security Policy. (2019). *Report on the implementation of the Action Plan Against Disinformation* (Joint communication 52019JC0012). Retrieved from https://eeas.europa.eu/sites/eeas/files/joint_report_on_disinformation.pdf

European Union. (2019). *Cybersecurity in Europe: stronger rules and better protection*. Retrieved from https://www.consilium.europa.eu/en/policies/cybersecurity/

Government of Finland. (2019). *Finnish Cyber Security Strategy*. 44.

Habermas, Jürgen. (1991). *The structural transformation of the public sphere: An inquiry into a category of bourgeois society*. Trans. by Thomas McCarthy. MIT Press.

Hansen, L., & Nissenbaum, H. (2009). Digital Disaster, Cyber Security, and the Copenhagen School. *International Studies Quarterly*, *53*(4), 1155–1175. Retrieved from JSTOR.

Kaiser, R. (2015). The birth of cyberwar. Political Geography. 46. 11-20. 10.1016/j.polgeo.2014.10.001.

Kukkola, J., Ristolainen, M., & Nikkarila, J.-P. (2017). *GAME CHANGER Structural transformation of cyberspace.* Riihimäki: Finnish Defence Research Agency

Lacy, M., & Prince, D. (2018). 'Securitization and the global politics of cybersecurity'. Global Discourse Vol.8(1), pp.100-115.

Lausuntopalvelu.fi. (2019). Retrieved November 2, 2019, from Lausuntopalvelu.fi website: https://www.lausuntopalvelu.fi/FI/Proposal/Participation?proposalId=56ab7d09-1859-4815-a19a-1ef624eea034

Lehto, M., Limnéll, J., Innola, E., Pöyhönen, J., Rusi, T., & Salminen, M. (2017). Finland's cyber security: the present state, vision and the actions needed to achieve the vision. Publications of the Government´s analysis, assessment and research activities 30/2017; Prime Minister´s Office. Retrieved from: https://tietokayttoon.fi/documents/10616/3866814/30_Suomen+kyberturvallisuuden+nykytila%2C+tavoitetila+ja+tarvittavat+toimenpiteet+tavoitetilan+saavuttamiseksi_.pdf/372d2fd4-5d11-4991-862c-c9ebfc2b3213/30_Suomen+kyberturvallisuuden+nykytila%2C+tavoitetila+ja+tarvittavat+toimenpiteet+tavoitetilan+saavuttamiseksi_.pdf?version=1.0

Mikkola, H., Aaltola, M., Wigell, M., Juntunen, T., & Vihma, A. (2018).

Hybridivaikuttaminen ja demokratian resilienssi. *FIIA - Finnish Institute of International*

*Affairs Report 55.* Retrieved from https://www.fiia.fi/wp-content/uploads/2018/05/fiia_report55_web_hybrdivaikuttaminen-ja-resilienssi.pdf

Ministry of Defence. (2010). *Security Strategy for Society*. Retrieved from "https://turvallisuuskomitea.fi/wp-content/uploads/2015/10/yts_2010_fi_nettiin.pdf"

Patomäki, H. (2015). 'Absenting the absence of future dangers and structural transformations in securitization theory'. International Relations, 128-136.

Peoples, C., & Vaughan-Williams, N. (2010*). Critical security studies: An introduction.* London: Routledge.

Perit, P., & Tuohy, E. (2013). 'Cyber Space in Estonia: Greater Security, Greater Challenges´ International Centre for Defence Studies.

Rantapelkonen, J., & Kantola, H. (2013). Insights into Cyberspace, Cyber Security, and Cyberwar in the Nordic Countries. *The fog of Cyber Defence*. 2013. 24 –36. Eds. Jari

Rantapelkonen, J., & Salminen, M. National Defence University: Helsinki.

Roe, P. (2008). Actor, Audience(s) and Emergency Measures: Securitization and the UK's Decision To Invade Iraq. Security Dialogue 39. 615-635.

Russell, A. L. (2014). Cyber blockades. Washington DC: Georgetown University Press.

Security Committee. (2019). Turvallisuuskomitea – toiminta ja tehtävät – Turvallisuuskomitea. Retrieved November 3, 2019, from https://turvallisuuskomitea.fi/turvallisuuskomitea/turvallisuuskomitea-toiminta-ja-tehtavat/

The Security Committee. (2013). *Finnish Cyber Security Strategy.* Retrieved from https://puolustusvoimat.fi/documents/2182700/0/Kyberturvallisuusstrategia/bb56d179-9b3a-4816-806d-84c84b04da30

The Security Committee. (2017). *The Security Strategy for Society*. Retrieved from https://turvallisuuskomitea.fi/wp-content/uploads/2018/04/YTS_2017_english.pdf

The Security Committee. (2019). *Finnish Cyber Security Strategy*. Retrieved from https://turvallisuuskomitea.fi/wp-content/uploads/2019/10/Kyberturvallisuusstrategia_A4_ENG_WEB_031019.pdf

The Security Committee. (2019). Operation and responsibilities. Retrieved from https://turvallisuuskomitea.fi/en/security-committee/the-security-committee-operation/.

Tuukkanen, T. 2013. Sovereignty in the Cyber Domain. *The fog of Cyber Defence*. 2013. 37–45 Eds. Jari Rantapelkonen & Mirva Salminen. National Defence University: Helsinki.

Veebel, V. & Ploom, I. (2016). Estonian Perceptions of Security: Not Only About Russia and the Refugees. *Journal on Baltic Security* 2-

Wæver, O. (2011). 'Politics, security, theory'. Security Dialogue Vol.42 pp.465-480.

Wigell, M. (2019). Democratic Deterrence: How to Dissuade Hybrid Interference. *FIIA – Finnish Institute of International Affairs Working Paper 110.* Retrieved from https://www.fiia.fi/wp-content/uploads/2019/09/wp110_democratic-deterrence.pdf

# Part II

# Elections

# 2.1 The Digital Revolution and Deliberative Democracy: A Tale of Two Referenda

Adam Oliver Smith, Juhani Mäntyranta, Jooel Heinonen, Anna Hattunen, Dominic O'Hagan
Faculty of Social Sciences, University of Helsinki

# Abstract

Referenda and digital communications platforms both offer an inherent promise of enhanced democratic participation. This paper will argue that the 2014 Scottish Independence Referendum illustrates how the relationship between digital communications and direct democracy can yield positive outcomes, whilst the subsequent "Brexit" referendum two years later suggests that such outcomes are no longer attainable, owing to shifts within the digital environment in which we all inhabit. It will conclude that, pending significant legal, social, and cultural changes regarding how we communicate online, it is no longer possible to conduct a referendum that meets normative standards for a deliberative democratic environment.

*Keywords*: Referenda, direct democracy, deliberative democracy, data-driven campaigning, public sphere, social media, propaganda, UK, United Kingdom, Brexit, Scottish Independence, IndyRef, EU, European Union

# Introduction

Whilst referenda in some form or another have always been a component of deliberative democracies, the use of them to put political issues to national electorates has intensified since the second half of the Twentieth Century, reaching an all-time high this decade (Mendez & German, 2016). However, recent prominent examples of referenda have been tainted with claims of illegality, unfairness, and undemocratic conduct. In the UK, which has held two prominent national referenda in the past five years, such claims were intrinsically linked to the use of digital technologies and communications. Namely, the use of personal data by public and private actors, the use of algorithms to "micro-target" voters with often misleading or false information, the role played by social media platforms, and the ability of campaigners to use technology to circumvent electoral law have all been highlighted to suggest that the 2016 "Brexit" referendum was illegitimate.

In this paper, we will compare and contrast two referenda held within the UK: the 2014 Referendum for Scottish Independence ("IndyRef") and the 2016 Referendum on the United Kingdom's Membership of the European Union ("Brexit"). These two referenda took place according to near-identical laws and regulations, and within a small timeframe.

There are a number of reasons why this paper focuses on direct democracy, as opposed to representative democratic exercises. For one, the promises that digital technology will widen participation are "the same promises made by the referendum instrument…in each case, voters are asked directly what should be done" (Floridi, 2016, p.189). In addition, referenda and digital communications tend to be framed as egalitarian, promising an unfiltered form of democratic and social participation that empowers everyone to participate in the public sphere. This paper will therefore examine the use of digital technology in both referenda to argue that, while the 2014 referendum represents the promise of direct democracy and digital communications, the 2016 referendum illustrates that a sea change in the way that digital communications are utilised for campaigning. This suggests that such ideals can no longer be realized, pending significant social and legal changes.

This paper consults a number of philosophical and sociological thinkers with regards to direct democracy and communications, namely John Rawls, Jürgen Habermas, John Searle, Luciano Floridi, and Plato. We also interviewed Adam Ramsay, the co-editor of Open Democracy, an investigative journalism organisation that has covered both referenda extensively, who played an instrumental role in guiding our research.

## A “fair” referendum — Anna Elina Hattunen

This section aims to provide a comprehensive definition of what may be considered a ‘fair’ and “good” referendum. In other words, the chapter discusses the key concepts surrounding referenda and whether they are a useful tool for measuring public opinion in a deliberative democracy. The aim is to provide an explanation of a “valid” and “fair” referendum which is at the same time the fundamental basis for this paper. Our main aim is to gain an understanding of whether Brexit and the Scottish independence referendum were “fair” and how the impact of digital technologies affects the implementation of a referendum in general, and in these two cases in particular.

A referendum has been a tool for governments and other institutions to solve problems and disagreements that governments, political parties and parliaments are, for several different reasons, unable or unwilling to solve (Gallagher et Al., 1996 p.1). In general, a referendum can give legitimacy for rather far-reaching changes (Henley et Al., 2019). A referendum refers to a mass electorate votes concerning a specific public issue, such as independence movements as well as other constitutional matters. It is worth mentioning that we are discussing referenda that are evolving around more complex topics such as independence, as is the case with Scotland, and the resignation of the United Kingdom from the European Union (Brexit). In western Europe, referenda are often seen as a tool for people to mobilise and participate actively in the decision-making process. They have been said to represent a legitimate decision straight from the people. In most cases, referenda have also increased political participation in general. First and foremost, it is a decision-making process that democracies more or less all around the world have used over the decades (Gallagher et Al, 1996 p.2).

Referenda have gained popularity within the past few decades, and since 1973 over 600 of them have been conducted around the world (Henley et Al., 2019). In the United Kingdom, referenda have particularly been a prominent part of political decision-making and political processes for more than 40 years now (Atkinson et Al, 2017 p.5). This is despite the fact that the UK has no constitutional requirement to hold a referendum when instigating constitutional change. Some authors and academics have discussed referenda as a tool to gain a wider perspective on a certain public opinion or issue. In other words, they have been used to perceive rather directly what public opinion on a certain issue or change is; they discern an answer or opinion from the people (Henley et Al., 2019). Referenda are often put forward to allow people to participate in a more detailed and determined way within the political decision-making process. This aims to foster greater levels of political participation among citizens (Gallagher et Al, 1996 p.2).

It has been claimed that referenda have been a great tool to engage people from marginalized and minority groups to participate in decision-making processes. These groups are in danger of being left out of the decision-making process due to their unwillingness, for whatever reason, to participate in the democratic process. This has been an alarming phenomenon for many policy makers in Western European countries.

John Rawls also emphasized the importance of the equal accessibility of information to everyone. This is a core requirement for a ‘fair’ society (Fallis, 2007 pp.23-36). As said, referenda are an attempt to shape and construct an understanding of the common or public will of the people. Therefore, one great feature for the definition of a “fair” referendum is to engage as many people as possible, in order to gain a credible view of a certain public opinion that is representative of the majority. Many theorists view referenda as a valuable part of democratic systems and governments. They are, indeed, part of a deliberative democratic system (Atkinson et Al, 2017 p.5)

The key principle and criterion for referenda in the United Kingdom is to try to clarify the public’s views and ensure that major constitutional changes take place with relatively broad consent. Another criterion for a “valid” and “fair” referendum is the fact that it is not desirable to create a circumstance in which the government of the United Kingdom feels forced to implement a policy that is actually against the will of the British Parliament (Ibid., p.8). Referenda, in the United Kingdom, are also not used in order to solve intra-governmental problems. There is an ongoing debate and discussion about whether there should be broader options and multi-option referenda (Ibid., p.9). In our case studies, however, this is not a valid question since we are looking at two rather similar referenda that did not have a multi-option agenda. When it comes to referenda, there are usually two types of results: referenda can be either legally binding and formally indicative, or serve as guidelines for the parliaments and different decision-making bodies and institutions (Gallagher et Al, 1996 p.12). Referenda are technically not binding in the United Kingdom since parliament is sovereign. As a criterion for a legitimate referendum in the United Kingdom, it is also important that the government or other actors are not able to manipulate the result of the voting (Atkinson et Al, 2017 p.9). Crucially, “[r]eferendums should never be regarded as producing mandates that override regular principles of representative democracy and the rule of law” (Ibid., p.9).

The most crucial point in defining a “fair” referendum is the engagement of people without falsifying and manipulating information through different channels. As mentioned previously, referenda can be a way to engage people, especially those from marginalized groups. Adam Quinn, from the University of Birmingham, has also stated the following: “The referendum campaigns themselves can be savagely divisive, especially when the prospect of a narrow victory tempts campaigners to use every argument at their disposal” (Birmingham.ca.uk, 2019). Although Quinn’s statement could hypothetically be applied to either IndyRef or Brexit, the later campaign was marked with particular hostility, such as the murder of an MP (BBC News, 2016). Namely, the use of personal data to affect people’s opinion and micro-targeting voters with misleading or outright false information are major concerns in terms of the referendum’s legitimacy. The data indicates that the Brexit referendum failed to meet the criteria for a “legitimate” and “fair” referendum.

One could argue that there are some criteria that need to be met for a referendum to be considered “valid” and “fair”. As mentioned, there are also differences between referenda and how they are conducted. In the United Kingdom, a “fair” referendum is an attempt to mobilise a large group of the population to vote and participate. However, the use of false and misleading information, as an example, could be argued to lead to the Brexit vote being considered unfair and possibly even illegitimate. Further on, we introduce our two case studies as well as discussion on the negative and positive aspects of a referendum. Our aim is to provide an understanding of the possibilities of the referenda in a fair context. Therefore, our aim is to gain a broad perception of the impact of digital technologies on liberal deliberative democracies by examining two different case studies.

## Case study I: The Scottish independence campaign, a model or a different age? — Dominic O’Hagan

The result of the 2014 Scottish Independence referendum was the rejection of independence, or the ‘No’ campaign, winning 55.3% of the vote compared to the Yes campaign’s 44.7% of the vote. With a turnout of 84.6%, the referendum represented the highest voter participation in any UK election or referendum since the introduction of universal suffrage (BBC, 2014). Although there have been some complaints the referendum was unnecessarily “divisive”, there has been general agreement that the referendum represents a good example of how a mass participatory referendum can re-vitalize democracy and can lead to a flourishing of debate (McWhirter, 2015 p.4).

The Scottish independence referendum’s regulatory framework was established by the Scottish Independence Referendum Act 2013 (Tickell, 2014 p.407). However, as Tickell points out much of the Act “drew heavily” from UK legislation, namely the Political Parties, Elections, and Referendums Act 2000 (PPERA). Because of this it can fairly be stated that both IndyRef and the Brexit vote were conducted in a near identical legal and regulatory environment.

However, even prior to IndyRef, some had pointed out that PPERA was already out of date, especially with regards to online communications and social media (Tickell, 2014 pp.407–408). In particular, the rules over the restriction of the publication during purdah (pre-election/referendum period) were outdated. Tickell points out that “publishers and promoters” includes social media users. At the same time, failure to follow strict electoral laws surrounding publishing could be considered a criminal offence (Ibid., pp.408–409).

Prior to the vote, the Electoral Commission pointed out that they would have a commonsense approach to what would be considered publication (Ibid., p.409). However, due to the fact that legislators and regulators had not taken into consideration developments in communication technology since the early 2000s, regulation was at best ad hoc and at worst totally unenforceable. Although this is a small example, it highlights the lack of forward planning and gaps that already existed in 2014.

This lack of foresight was particularly glaring due to the vibrant social media environment that IndyRef took place in. In 2012 British Telecom found that Scottish households used social media more than another part of the UK, with 48.2% making use of these platforms (Ibid., p.406). IndyRef brought social media engagement into the political process like never before. In the 24-hour period during polling day there were 2.6 million tweets relating to Indyref. In the five weeks prior, there were ten million Facebook interactions (McNair, 2015).

Despite social media being utilized widely, its use was not uniform across the debate. Many studies have shown that Yes Scotland far outperformed the Better Together campaign in terms of engagement and driving the online debate (Shephard, Quinlan 2015 p.481; Langer, Comerford, McNulty 2019, p.846). This is explained via a number of factors, including: the average Yes supporter was younger, the lack of mainstream media support for independence driving Yes supporters online, and the structure of the campaigns themselves (Ibid, p.847).

During the IndyRef campaign, not a single daily newspaper in Scotland came out in support of Yes. Only one mainstream media outlet, the *Sunday Herald* declared for Yes (Press Association, 2014). McNair (2015), points out that due to the lop-sided nature of the debate within traditional media outlets, "social media provided space for subjectivity, opinion and overt ideological bias to be expressed". While it could be contested that the same "ideological bias" exists within the mainstream media, it is clear that social media was seen as a way at redressing the imbalance.

A study looking into the internet usage of people searching for information on the IndyRef found that people valued "facts" more than opinion (Baxter, Marcella 2017 p.8). Despite this, the study also found that, "… social media messages (not necessarily from official campaign groups and parties) were more powerful than fairly static websites in dominating political discourse and reaching a wider swathe of the voting public" (Ibid., p.15). This would seem to indicate a level of trust in social media, at least in 2014, compared to official sources.

The differences between Yes Scotland and Better Together, the two official campaign groups, also played a significant role in how social media was used. Better Together very much carried out a "top-down centralized command and control" campaign (Langer et Al., 2019 p.836). On the other hand, Yes Scotland committed itself to a more "hybrid model" that, "… had elements of a more decentralized structure and some movement-like dynamics blended with traditional party campaigning characteristics, enabling—or at least not discouraging—more autonomous bottom-up participation and entrepreneurial modes of engagement" (Ibid., p.846). It has been argued that this de-centralised strategy not only allowed for free-flowing debate, but actually enabled more grassroots campaigning.

Some within Yes Scotland's campaign team have stated their belief that "… social media played an important, and often crucial, role in the formation of many of the grassroots groups that emerged during the campaign", citing National Collective as an example (Ibid., p.848). These groups followed a pattern of online discussion: online organising, and offline activity. This indicates that social media had an overwhelmingly positive effect in engaging people in the democratic process, beyond being mere consumers of information.

The social media-engaged Yes supporters were far more likely to form independent groups. Yes Scotland had 350 affiliated local groups compared to Better Together's 80. Yes Scotland were also able to lean on groups and individual bloggers, or 'online allies', creating content that more or less did not exist for Better Together (Ibid., p.847). In an interview for this research paper Adam Ramsay, Editor of OpenDemocracy, made the suggestion that the grassroots nature of IndyRef may have protected it from some of the negative interference and allegations of illegality that have plagued the Brexit campaign: "Grassroots campaigns tend to be better than the 'astroturf' campaigns run by distant campaign offices" (Ramsay, 2019).

Although there may be some validity to that argument, it does not address the fact that the IndyRef campaign was also conducted in a radically different digital environment compared to Brexit or even in the elections in the following years. Some of the reporting of social media's influence in IndyRef can at times seem quaint compared to 2019. Rather than debates surrounding micro-targeting, data harvesting, and so called "dark money", the discussion in 2014 focused instead on 'likes' and 'Facebook engagement' (Riddell, 2014).

The growth in microtargeting can be seen when evaluating how much political parties spent on Facebook ads. In the 2015 UK-wide general election the three biggest parties in Scotland (SNP, Conservatives, and Labour) combined spent £42,121 on Facebook ads. However, just one year later in the 2016 Scottish Holyrood election '… overall spending on Facebook ads increased by a third in Scotland with Labour alone nearly trebling their outlay' (Ellison, 2017).

It should be noted that although IndyRef is seen as a model in how referenda can increase democratic participation and how they can be used to negotiate complex problems, it has to be seen in the appropriate context. The Scottish referendum resulted in a status quo outcome, whereas, Brexit is a radical departure from the status quo. Due to this, allegations of deception and misleading "facts" may be more in-focus because of the dramatic changes that Brexit will bring about.

## Case study II: The Brexit referendum campaign, social media & divergent realities — Adam Oliver Smith

The 2016 'Brexit' referendum, the result of which was a "leave" vote that is currently causing convulsions in the British body politic, is an illustrative example of how the digital revolution is testing the limits of deliberative democracy. As explained earlier, the 2016 referendum took place in a broadly similar social and legal environment to the 2014 referendum on Scottish Independence. However, the Scottish Referendum has not been subject to the level debate, analysis, and scrutiny regarding the role of digital misinformation that the Brexit campaign has been subjected to. This is because the two referenda occurred in separate and distinct digital environments.

The use of automated social media bots to spread misinformation, the deployment of psychometric-driven propaganda by the Leave campaign, and the use of big data processes to curate and deploy "microtargeted" messages to swing voters are activities that have been afforded ample attention within the rich outpouring of academic research on the 2016 referendum. To draw on just one example, computational analysis of social media networks that began hours after the result was announced revealed that 14,000 Twitter bots published 1.5 million tweets (75% of which were identified as pro-Leave messages) in the week before polling day, before promptly disappearing from the platform altogether (Bastos and Mercea, 2018 pp.1-2). In addition, data-driven campaigning firms in the employ of Vote Leave, including the now-defunct Cambridge Analytica, utilized the personal data of hundreds of thousands of British voters to send millions of microtargeted, sometimes contradictory political advertisements to different segments of the electorate in their efforts to produce a Leave majority at the polls (Bay, 2018, pp.1-14). While the effectiveness of these tactics in influencing the result are debatable, it is evident that the UK electorate was subjected to an intensive, population-scale misinformation campaign that was enabled and facilitated by the digital communications platforms that are now an implacable part of our daily lives.

Before diving into the philosophical concerns surrounding the Brexit campaign and its implications for deliberative democracy, it is important to clarify what makes this case study distinct from the Scottish Independence referendum. There is a general consensus that the scale and nature of digital misinformation deployed in the Brexit campaign was, at the time, unparalleled in British political history. The use of psychometrics and data-driven microtargeting first came to the attention of strategists in the UK following the publication of a 2014 research paper outlining the effectiveness of "behavioural targeting" in influencing the real-world behaviour of social media users (Chen and Stallaert, 2014 pp.429-449). It was this research, published after the 2014 Scottish Referendum had concluded, which was cited by Vote Leave front man Dominic Cummings as playing a central role when informing the Leave campaign's digital strategy (Cummings, 2017). Taken in conjunction with the scant amount of evidence suggesting that campaign groups on either side of the Scottish referendum utilized such tactics, we can reasonably assume that the two referenda took place in wholly separate digital environments.

The goal here is not to debate the effectiveness of social media bots and microtargeting in referendum results. As has already been observed, it "strains credulity that… [the success of] the Leave campaign was entirely the product of manipulation… rather than the surprising success of a political campaign that tapped into the attitudes and beliefs held by millions of people in the United Kingdom" (Benkler et al, 2018 p.341). Rather, the goal is to outline the ways in which the pervasiveness of such tactics has shown that the Brexit referendum failed to meet the standards for a functional public sphere and healthy democratic environment. The purpose is to impress upon the reader the seriousness of the challenge that lies before us. If measures are not taken to curtail, regulate, and provide accountability, then deliberative democracy will cease to be a viable system.

As Benkler et al. set out in their landmark analysis of "network propaganda", a population with vastly divergent, competing conceptions of reality can never be stable, nor can it sustain a functional democracy (Ibid. p.5). A healthy public sphere requires the participants to at least share the same social and moral universe as each other, albeit one where viewpoints may diverge considerably (Habermas, 1984). However, when uninhibited, aggressively targeted misinformation campaigns can be conducted on a population-wide scale, obstacles begin to appear that inhibit the ability of a society to construct and exist within a consistent shared social reality. Take for example the use of social media psychometrics to deliver tailor-made misinformation to individual users, a strategy which Vote Leave campaigners continue to publicly cite as critical to their success (Conoscenti, 2018 p.71). In the weeks before the 2016 referendum, millions of people received contradictory messages about the EU which were sent to them based on what the data said would trigger the desired response. The result is that different segments of the population have disparate and disconnected conceptions of what reality in the UK and EU actually looks like. This is a dangerous situation for an already polarized society to be in.

As Habermas points out, any sort of discourse in the public sphere must be held to certain norms of truthfulness in order for it to benefit democracy (Habermas, 1990, p.35). In addition, he illustrates the need for a shared language, context, and morality system in order to test validity claims and deliberate collectively (Habermas, 1984, p.10). The ability (and intention) of psychometric-driven campaigning to effectively “crowd out” alternative sources of information is an inherent threat to these ideals and structures. Furthermore, if we are to take the position that metaphors, and the speech acts of which they constitute a central component “have a massive influence on the construction of reality” (Walter and Helmig, 2008 p.119), then the threat posed by psychometrics and microtargeting by the Leave campaign takes on another dimension.

The success of the Leave campaign was largely built on its ability to use technology to construct the EU as a metaphor for public anxieties around immigration and inequality, as countless frame analyses of their political advertisements have concluded (Conoscenti, 2018 pp.65-82). Their success suggests that the emergent digital technologies used in the Brexit campaign have the capacity to fuel different constructions of social reality within a single electorate. To take this point further, one should also take Rawlsian ethics into account, particularly with regards to equity of information access to inform rational decision-making in a democracy. Previous analyses of Rawls in relation to data ethics have concluded that the construction of the “fair society” requires the public to have equal access to information, particularly when making democratic decisions (Fallis, 2007 pp.23-36). One would struggle to describe a public that is consistently bombarded with personalised misinformation as enjoying sufficient, equal access to the information needed to make direct democratic decisions.

Although it is doubtful that the digital strategies deployed by the Leave campaign were the primary cause of the referendum result, their undeniable efficacy raises broader concerns about the sustainability of both direct and deliberative democracy. The analyses provided here should highlight the urgency with which psychometrics and data-driven misinformation campaigns must be tacked, regulated, and held to account if we ever wish to conduct a ‘fair’ referendum in the future.

## The threat — Juhani Mäntyranta

This chapter will look at arguments against the use of referenda in complex, multi-issue topics. First, representational democracy is examined as the main democratic alternative to referendum, after which Platonic and Habermasian views on how a functional democracy operates are examined. These two topics also constitute the main argument as to why a referendum is not necessarily the best choice for deciding complex issues. Further, the ongoing digital revolution and its effects on political deliberation are examined with regards to referenda. The topics examined are the changing habits in news consumption, the spreading of misinformation, and the spreading of disinformation.

Most Western democracies are representative in nature. What that means is that the public will vote for their selected candidate in an election, to represent them in parliament, congress, or executive office. The reason for the representational nature of the political system, in essence, is because “direct deliberation among all citizens is widely assumed to be impossible on the scale of the modern nation state” (Landemore, 2017, p. 2). Because of the complex nature of the issues debated, it makes sense that career politicians, ones that are able to examine the issues from multiple angles and through different lenses, are in charge of the decision-making process. Most non-politicians will not be able to make the effort of examining said issues as extensively as elected members of parliament can. Thus, the ability of most non-politicians to form a well-rounded understanding of issues can be limited.

The previous chapter also ties in with both the Platonic, as well as the Habemasian theories of democracy. Plato, speaking at a time of direct democracy, argued that (direct) democracy can only work if the public is educated and well informed on the topics they are voting on. He also argued that most people are not educated enough to be able to partake in the democratic process and thus proposed a system of technocratic governance, in which the ruling class has been educated in the realm of governance from a young age (Plato, 380 B.C./1965, Republic VI). Habermas’ ideas on deliberative democracy echo some of the same sentiments. According to Habermas, the public sphere is a place where people should be free to discuss and debate all things related to politics (Habermas, 1984). A well-functioning public sphere is a prerequisite of deliberative democracy. However, when the public sphere is filled with polarisation, it is much more difficult for the public to be able to debate issues. Because it is more difficult to debate issues, direct democracy, in all its forms, becomes a less than ideal way of making decisions, especially on complex issues.

One of the vexing issues of modern democracy, as mentioned above, has to do with polarisation. As debate on political issues has shifted more and more towards the online sphere, one issue that has arisen is the increased amount of polarisation and partisanship (Pattie & Johnston, 2016, p. 484). This is a huge problem for deliberative democracy, as people will have lower levels of engagement with people of alternative views and opinions and thus, will only be partially informed on matters. In a representative democracy, the situation explained above would be a problem, but not as much as in a direct democracy, or one with referenda on complex issues. The reasoning is that in a representative democracy, politicians will be more informed, and there is an expectation that they will at least try to supply the required information on any given topic, as well as debate topics within parliament. In a direct democracy, if issues between people of opposing views are not debated, then deliberation is taken out of the equation. Without deliberation, direct democracy will not be effective enough to bring about decisions that are based on reason and deliberation.

Another issue that creates problems has to do with new ways in which the news is consumed. Traditionally, people would get their news from newspapers or television. There were variances between individuals on how much exposure they got, especially with changes in the media landscape (Shehata & Strömbäck, 2011, pp. 111). What can be argued is that those individuals who opted to watch the news on TV, as an example, were more likely to watch the news from start to finish. Revealingly, a recent study suggests that in the modern social media ecosystem, news is not read in full, but rather only the headlines and previews of articles are taken in (Anspach, Jennings, and Arceneaux, 2019, pp.1-2). Indeed, the authors claim that even though 68% of Facebook users use the site for news, an average user only clicks on 7% of the political news stories in their feed, which in turn leads to audiences' overconfidence in their knowledge of the issues (Ibid. p.2). The results of the study found that people who only read the preview were less informed on issues and were less likely to answer questions on the topic correctly, when compared to those who read the entire article. The findings suggest that people might be overconfident about their understanding of different topics and possibly vote on issues with limited or incorrect knowledge.

One major issue with holding referenda on complex issues has to do with the internet at large. The amount of fake news, disinformation and inaccurate information has grown in recent times. In a study by Vargo, Guo & Amazeen, (2018) the authors find that the amount of fake news increased in the United States between the years 2014 and 2016 and it is very likely to have increased even further in the subsequent years. We can all remember how one of the Brexit leaders was posing next to a bus with an advert stating that the £350M the United Kingdom (UK) pays the European Union (EU) on a weekly basis could be used to fund the UK's healthcare system. Boris Johnson later backtracked on this claim, but the message had already been sent and the correction of the disinformation cannot change the result of the referendum (Read, 2019). Claims such as the promise of £350M can spread like wildfire within the current social media ecosystem, giving voters misinformation on different subjects. Indeed, the same can be seen to be the case with normal elections as well, but the members of the parliament will be more informed on issues and are thus less likely to believe false claims.

It can also be argued that non-politicians are more easily swayed, not only by public influencers, such as politicians, but also by automated social media accounts that spread disinformation, also known as bots. These bots function by automatically spreading misinformation, as soon as an article has been released, in order for the article to gain exposure, as well as to gain online legitimacy, by having a large amount of shares (Shao et.al., 2017, pp.1-3). By spreading said articles, it is easy to manipulate ill-informed members of the public into believing untrue statements, as well as to then base their vote on said statements. Indeed, the Leave side of the Brexit referendum utilised such tactics to great success (Bastos and Mercea, 2018, pp.1-2). It is easy to argue that the usage of bots can be utilised in a negative way, in democratic societies in general, but especially with regards to referenda.

## The opportunity — Jooel Jacob Heinonen

Even though the fourth digital revolution may have weakened the ability of liberal democracies to hold fair and genuinely democratic referenda, new methods of digital communications also have the potential to enhance and increase democratic engagement by citizens and facilitate greater direct participation. This section discusses the positive effects digital communications can have on the ability to strengthen deliberative democracy through direct participation and identifies the major opportunities granted by the digital era to create a more vibrant, inclusionary and participatory democracy.

Deliberative democracy is a form of direct democracy that is based on rational debate and collective decision-making. The ideal of deliberative democracy is to reach rational, consensual decisions while being a more inclusionary and participatory model of democracy. Deliberative democracy would ideally facilitate greater direct participation by citizens. The core of deliberative democracy is the idea that authentic deliberation would vest policies with greater legitimacy, acceptance and quality than would the strictly representative modes of decision-making (Friess and Eilders, 2015, p. 319). The dawn of the digital era can be seen as an opportunity to create and foster a truly inclusionary and participatory arena of debate and discussion in which the ideals of deliberative democracy can truly be realized. The virtual space created by the internet provides, for the first time, the ideal conditions for deliberative democracy by offering an infrastructure for the public sphere that the theorists of deliberative democracy have only been able to dream about so far (Ibid., p.320). The internet and new digital communications have the potential, therefore, to contribute to the creation of a stronger deliberative democracy.

The internet provides new venues for debate and discussion and offers ways to create a more inclusionary and participatory public sphere, where deliberation takes place. The digital public sphere and the new media (social media, online forums, online news sites) are created through the participation of the many, while still exhibiting the features of an enlightened public: free access, no restrictions, no limit on the number of participants, and equality of access among the participants (Kreide, 2016, p.481). Unlike the old mass media, the new media does not distinguish between an audience and the public, and the usage of the media is designed through the users, not just for them (Ibid., p.481). Ideals of a well-functioning public sphere can be realized through interactive internet communities that are based on forms of joint participation by members (Ibid., p.481). The internet also has the potential to bring new voices into the public sphere. People who feel disinclined to engage in face-to-face political conversations may be more willing to do so in online spaces. Indeed, according to a study by Jennifer Stromer-Galley (2014), political conversation with strangers and acquaintances is considered by many people a taboo activity, and the internet may offer a new context for political conversations for those that are unwilling to engage in face-to-face political conversations. This would introduce more people to the public sphere, where they can engage in deliberation on societally significant issues (Stromer-Galley, 2014). A more participatory and inclusionary public sphere, enabled by technological progress and digital communications, can increase the voice of the public in decision-making and therefore, since authentic deliberation is the source of the legitimacy of the law in deliberative democratic theory, strengthen the legitimacy of major political decisions.

In addition to contributing to the creation of a more participatory and inclusionary model of deliberative democracy, the digital revolution and the internet also have the potential to positively influence the quality of political debates and discussion through an equal access to information, as Antje Gimmler (2001) shows in her study on the impact of the internet on the public sphere and deliberative democracy. Equality of access to information, as well as the access being unrestricted, are fundamental preconditions for the creation of a more ambitious practice of discourse (Gimmler, 2001, p.31). The internet can be used to support these aims, since the internet encourages the exchange of information by making it easily obtainable to users at a very low cost (Ibid., p.32). As a result, the ideal of the deliberative model of democracy, where well-informed and rational citizens make decisions through careful deliberation, could be realized through the possibilities offered by the internet and equal access to information. However, recent developments have illustrated that equal access to the internet for all citizens is not guaranteed, as the Brexit experience demonstrates. Nevertheless, the internet has the potential to create a more informed citizenry through the availability of information online, and thus increase the quality of deliberation, which is essential for the creation of a more meaningful model of deliberative democracy that is based on rational discussion.

Typically, the deliberative model of democracy would involve citizen participation at every stage of the political process, whereas a referendum only brings citizens to the decision-making process at the very end. Referenda and the deliberative model of democracy seem to be somewhat at odds with each other, since a referendum takes place to settle a particular question at the ballot box, whereas the deliberative model is more interested in the process of reaching a decision through rational and well-informed discussion and debate. However, Lawrence LeDuc (2015) argues that there are ways to revise the familiar institutions of initiatives and referenda to approximate a more deliberative form of direct democracy. LeDuc concludes that direct democracy can become more deliberative in practice if the overtly partisan motives behind several referendum campaigns can be controlled, if better question wording and improved availability of information can lead to greater clarity on the issue of the referendum, and if citizens can be more engaged, resulting in more inclusive rates of participation (LeDuc, 2015). The participation and the engagement of the public in decision-making are essential requirements for the creation of a truly deliberative model of direct democracy. However, a high turnout in a referendum should not be confused with engagement and participation in a truly deliberative process (Ibid., p.146). In addition, the Brexit referendum demonstrates that engagement can be problematic in itself when fueled by disinformation. As an analysis of the online deliberation on the issue of the Scottish independence shows, the deliberative process can be quite narrow, even though the turnout in the Scottish independence referendum was very high (Ibid., p.146).

As discussed above, technological development and the opportunities offered by the internet can facilitate the creation of a more inclusionary and participatory model of deliberative democracy. As noted by LeDuc, a direct democracy requires a well-informed public and a high degree of participation by the public to become more deliberative in practice. As mentioned earlier, the internet and new digital communications have the potential to create a more participatory and inclusionary public sphere, where deliberation would take place. This could potentially realize the ideals of deliberative democracy; of citizens coming together to make decisions through rational discussion. New, emerging digital communications (social media, online forums, online news sites) enable the creation of a more participatory public sphere, where the measures of direct democracy, such as referenda, can be made more deliberative, instead of just being processes where the public's participation is limited to simply voting.

## Conclusion

This paper has attempted to look at the issues that surround referenda in the age of mass online communication and social media. To highlight some of the opportunities and dangers this paper has focused on the Scottish Independence referendum of 2014 and the UK's EU referendum of 2016.

Within the UK, referenda have only ever been utilized to gauge public opinion on complex constitutional issues. There have been 13 referendums in the UK since the first one in 1973, however, only three of these were UK-wide referendums (UK Parliament Website, Accessed November 2019). It is perhaps indicative of the current problems that surround Brexit that of those three UK-wide referendums, two have been on the issue of the UK's membership of the European Communities and later the European Union.

In one chapter of this paper the argument was made that referenda, as they are held in the UK, are too complex for ordinary people to fully grasp. However, many successful referenda in the UK dealt with equally complex issues to Brexit, not least

IndyRef. It is pointed out in the following chapter that referenda as a form of deliberative democracy can infer legitimacy on a complex topic. This paper takes the position that although referenda are messy and imperfect, the alternative of a quasi-technocratic and purely representative democracy is at least as problematic. This paper therefore does not take the position that ‘fair’ referenda are not possible in the era of online communication.

However, this position is not without caveats. As has been pointed out throughout this paper a referendum must be carried out within a regulatory structure that reflects the age and environment in which the vote takes place. These regulatory structures should encourage the development of a public sphere based on updated Habermasian ideals (see e.g. Dryzek 2008) and should promote the equality and parity of information in the model of Rawlsian theory. This paper has taken the position that due to the reasons outlined, the Brexit vote did not meet these criteria.

That is not to say that in our contemporary age referenda must always be viewed through a purely critical lens. The use of social media and online communication generally led to a flourishing of debate during the Scottish independence referendum. However, legislation and regulation, such as PPERA, have not kept apace to the changing face of online communication and campaigning.

Despite this, political parties and the government have shown little willingness in updating or replacing PPERA or introducing new additional legislation and projection. This is evidenced by the fact that the Electoral Commission’s reports have not made any mention of the effect social media, online misinformation, and microtargeting had on the EU referendum (Electoral Commission 2019). It would not be an exaggeration to say that the government has largely stepped out of the debate surrounding the regulation of referenda.

We find that referenda are neither inherently good or inherently bad. The same can be said for social media. However, if a flourishing public sphere is to be created, the rules that govern our modern body politic must be updated to meet the challenges posed by twenty-first century communication and campaign strategies.

## References

Anspach, N. M., Jennings, J. T., Arceneaux, K. (2019) 'A Little Bit of Knowledge: Facebooks News Feed and Self-Perceptions of Knowledge', *Research and Politics*, *6* (1) pp. 1-9.

Atkinson L, Blick A (2017) 'Referendums and The Constitution' The Constitution Society.

Bastos, M. & Mercea, D. (2018), 'The public accountability of social media platforms: lessons from a study on bots and trolls in the Brexit campaign', *Philosophical Transactions A*, *376* (2128), pp.1-12.

Baxter, G & Marcella, R. 2017. 'Voters' online information behaviour and response to campaign content during the Scottish referendum on independence.' *International Journal of information management* [online], *37*(6), 539-546.

Bay, M. (2018), 'Social Media Ethics: A Rawlsian Approach to Hypertargeting and Psychometrics in Political and Commercial Campaigns', *ACM Transactions on Social Computing*, *1* (4), pp.1-14.

BBC News. (2016). 'Jo Cox MP dead after shooting attack'. https://www.bbc.com/news/uk-england-36550304 [Accessed: 31/10/19].

BBC. 2014. Scotland Decides. https://www.bbc.com/news/events/scotland-decides/results [Accessed: 19/10/19].

Benkler, Y., Faris, R., & Roberts, H. (2018), *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics*, Oxford: Oxford University Press.

Birmingham.ac.uk. (2019). "*Referendums: the pros and cons - University of Birmingham*. [online] Available at: https://www.birmingham.ac.uk/research/perspective/referendums.aspx [Accessed 30 Oct. 2019].

Chen, J. & Stallaert, J. (2014), 'An economic analysis of online advertising using behavioural targeting', *MIS Quarterly*, *38* (2), pp.429-444.

Conoscenti, M. (2018), 'Big Data, Small Data, Broken Windows and Fear Discourse: Brexit, the EU and the Majority Illusion', *De Europa*, *1* (2), pp.65-82.

Cummings, D. (2017), 'On the referendum #21: Branching histories of the 2016 referendum and 'the frogs before the storm', *Dominic Cumming's Blog*, 9 January. Available at: https://dominiccummings.com/2017/01/09/on-the-referendum-21-branching-histories-of-the-2016-referendum-and-the-frogs-before-the-storm-2/ [Accessed: 29.10.19].

Dryzek, J. S., & Niemeyer, S. (2008). Discursive representation. American political science review, 102(4), 481–493. doi:10.1017/s0003055408080325

Electoral Commission (2019) 'Report: The regulation of campaigners at the referendum on the UK's membership of the European Union held on 23 June 2016', https://www.electoralcommission.org.uk/who-we-are-and-what-we-do/elections-and-referendums/past-elections-and-referendums/eu-referendum/report-regulation-campaigners-referendum-uks-membership-european-union-held-23-june-2016#footnoteref1_pl9twz1 [Accessed 31. October. 2019].

Ellison, M. 2017. ‘Election 2017: Scottish voters targeted by 'dark ads' on Facebook.’ *BBC News*. https://www.bbc.com/news/uk-scotland-scotland-politics-40170166 [Accessed: 31/10/19].

Fallis, D. (2007). ‘Information ethics for twenty-first century library professionals’, *Library Hi Tech*, *25* (1), pp.23-36.

Floridi, L. (2016), ‘Technology and Democracy: Three Lessons from Brexit’, *Philosophy and Technology*, *29* (3), pp.189-193.

Friess, Dennis, and Eilders, Christiane (2015). A systematic review of online deliberation research. *Policy & Internet, 7* (3), 319-339.

Gallagher, M. and Uleri, P. (1996). ‘*The referendum experience in Europe’*. New York: St. Martin's Press.

Gimmler, Antje (2001). ‘Deliberative democracy, the public sphere and the Internet’, *Philosophy and social criticism, 27* (4), pp- 21-39.

Habermas, J. (1990) *Moral Consciousness and Communicative Action*, Cambridge: Polity Press.

Habermas, J, (1984). *The Theory of Communicative Action Volume 1: Reason and the Rationalization of Society*, Boston: Beacon Press.

Henley, J., Carroll, R. and Rice-Oxley, M. (2019). ‘*Referendums: who holds them, why, and are they always a dog's Brexit?*.’ [online] the Guardian. Available at: https://www.theguardian.com/news/2019/mar/11/referendums-who-holds-them-why-and-are-they-always-a-dogs-brexit [Accessed 31 Oct. 2019].

Kreide, Regina (2016). ‘Digital spaces, public spaces and communicative power: In defense of deliberative democracy’, *Philosophy and Social Criticism, 42* (5), pp.476-486.

Landemore, H. (2017). ‘Deliberative Democracy as Open, Not (Just) Representative Democracy’ *Daedalus, 146* (3) pp. 1-13.

Langer, A I, Comerford, M, and McNulty, D. (2019). ‘Online Allies and Tricky Freelancers: Understanding the Differences in the Role of Social Media in the Campaigns for the Scottish Independence Referendum’. *Political Studies, 67* (4), pp.834-854.

LeDuc, Lawrence (2015). Referendums and deliberative democracy. *Electoral Studies*, *38* (1), pp.139-148.

MacWhirter, I. (2015). ‘*Tsunami: Scotland's Democratic Revolution.* Glasgow: Freight Books’.

McNair, B. (2015). ‘#indyref: the Scottish media and the independence referendum.*’ The Conversation.* https://theconversation.com/indyref-the-scottish-media-and-the-independence-referendum-37584 [Accessed: 31/10/19].

Mendez, F. and Germann, M. (2016). ‘Contested Sovereignty: Mapping Referendums on Sovereignty over Time and Space’ *British Journal of Political Science*, *48* (1), pp.141-165.

Parliament Website, (2019). ‘Referendums held in the UK’, https://www.parliament.uk/get-involved/elections/referendums-held-in-the-uk/ [Accessed 31. October. 2019].

Pattie, C., Johnston, R. (2016). 'Talking With one Voice? Conversation Networks and Political Polarisation' The British Journal of Politics and International Relations, *18* (2), pp. 482-497.

Plato. and Allan, D. (1965). *Republic*. London: Methuen. 484a-511e.

Press Association (2014). 'Sunday Herald declares 'yes' for Scottish independence.' *The Guardian,* https://www.theguardian.com/politics/2014/may/04/sunday-herald-declares-yes-for-scottish-independence [Accessed: 29/10/19].

Ramsay, A. (2019). Interview with Adam Ramsay. *Interviewed by O'Hagan, D & Smith, A. Skype Interview.* Data file and transcript available upon request. Interview date: 24/10/19.

Read, J (2019). 'Boris Johnson Appears to Finally Admit his '£350M a Week' claim Was Wrong' The New European https://www.theneweuropean.co.uk/top-stories/boris-johnson-350-million-a-week-nhs-claim-1-6264572 [Accessed 3.November 2019].

Riddell, J. (2014)- 'Scottish independence: how Facebook could change it all'. *The Guardian,* https://www.theguardian.com/media-network/media-network-blog/2014/sep/17/scottish-independence-referendum-facebook-social-media [Accessed: 31/10/19].

Shao, C., Ciampaglia, G. L., Varol, O., Flammini, A., Menczer, F., (2017) 'The Spread of Fake News by Social Bots', *Nature Communications, 9* (4787), pp.1-9.

Shehata, A. & Strömbäck, J. (2011). 'A Matter of Context: A Comparative Study of Media Environments and News Consumption Gaps in Europe', *Political Communication, 28* (1), pp. 110-134.

Sromer-Galley, J. (2014). 'New voices in the public sphere: A comparative analysis of interpersonal and online political talk', *Journal of the European Institute for Communication and Culture*, *9* (2), pp. 23-41.

Tickell, A. (2014). 'Regulating #indyref: Social Media and the Scottish Independence Referendum Act 2013'. *Edinburgh Law Review,* 18, pp.406-410.

Vargo, C. J., Guo, L., Amazeen, M. A. (2018). 'The Agenda Setting Power of Fake News: A Big Data Analysis of the Online Media Landscape from 2014 to 2016', *New Media and Society, 20*(5) pp. 2028-2049.

Walter, J. & Helmig, J. (2008). 'Discursive Metaphor Analysis: (De)Construction(s) of Europe' in Carver, T. & Jernej P. (eds.) *Political Language and Metaphor: Interpreting and Changing the World*, London: Routledge, pp.119-131.

# 2.2 How the Public Sphere has Changed Through Social Media and New Media Logics
# Case Study: The 2016 US Presidential Election

Laura Trémouille, Sonja Siirtola, Iiris Meurman, Mari Liukkonen,
Christa-Jemina Korhonen
Faculty of Social Sciences, University of Helsinki

## Abstract

Social media has changed the field of media and communications by providing ordinary citizens unlimited access to, and an ability to produce, circle and consume media content effortlessly. At the same time, it has become one of the central concepts of the communication research field and the arena of public debate. By using social media platforms such as Facebook and Twitter, users provide details of their personal life and interests to social media companies. Many of these companies use the provided data to provide advertisers a direct means to reach their target audiences. This has provided marketers with an opportunity to generate content to narrowly targeted interest groups, and this practice has been used widely by companies as well as political organizations and officials. Social media has also changed the logic of media and how the revenue in the media business is generated. This study discusses how social media and its algorithms have changed the public sphere by looking at the 2016 US presidential election, which has been called a "social media election". We also examine the impact new media logic has had on the public sphere by explaining how it affected the 2016 election. The key argument of this study is that through curated and targeted content, social media and the new media logics have changed the public sphere, and the effects were particularly well seen in the U.S. 2016 presidential election.

*Keywords*: Public sphere, social media, election, algorithms, targeted marketing, fake news, US presidential election 2016, Facebook, Twitter

## 1. The Age of Social Media

Social media platforms such as Facebook and Twitter provide a new communication forum where public figures can communicate widely to their audiences. This was particularly clear in the 2016 US presidential election, where both Donald Trump and Hillary Clinton made extensive use of social media in their election campaigns. For this reason, we have chosen to explore how social media has changed the public sphere through a case study of the US 2016 presidential election.

Social media has changed the field of media and communications by providing ordinary citizens unlimited access, and an ability to produce, circulate and consume media content effortlessly. At the same time, it has become one of the central concepts of the communication research field and the arena of public debate.

By using social media platforms such as Facebook and Twitter, users provide details of their personal life and interests to social media companies. Many of these companies use this data to provide advertisers with a direct passage to reach their target audiences. This has provided marketers an opportunity to generate content to narrowly targeted interest groups, and this practice has been used widely by companies as well as political organizations and officials. Social media has also changed the logic of media and how the revenue in the media business is generated.

This study examines how social media has challenged the position of traditional media, as publications of individual publishers can achieve much wider visibility than most popular traditional media outlets. How has this affected the public sphere, and what are the things that have contributed to this? The case study of the 2016 presidential elections shows how Donald Trump's tweets and fake news garnered more visibility during the election than social media publications of the most popular newspapers.

This study also discusses how social media algorithms have changed the public sphere. In addition, it examines the impact that new media logic has had on the public sphere by explaining how it affected the 2016 election. The key argument of this study is that through curated and targeted content, social media and the new media logics have changed the public sphere, and the effects were particularly clear in the U.S. 2016 presidential election.

Section two opens the case of the 2016 presidential election and explains why social media mattered in that election. Section three describes the concept of the public sphere and shows how it has changed with the development of media.

Section four explains how social media algorithms work and what effect they have had on the public sphere. Section five analyzes how targeted marketing can shape the public sphere and discussions. Section six explains the new media logic in the era of social media and describes how that affected the 2016 election and the public sphere.

## 2. Case Introduction – I. Meurman

Social media refers to online community services such as Twitter, Facebook, YouTube, and various blogs where users mostly produce their content on a non-commercial basis (Seppänen & Väliverronen, 2013, 36). This study discusses how social media has changed the public sphere and the role of social media in the 2016 US presidential election.

Although social media can be seen as a relatively recent phenomenon, already in 1993 Howard Rheingold popularized the idea of virtual communities, as he saw there an opportunity to create new communities that could reinvigorate public conversation and debate. For example, online social communities were seen as an opportunity to create a new space for social interaction and democratic participation (Hjorth & Hinton, 2019, 16). The rise of social media has raised questions about the nature of virtual communities again, because networked publics can emerge organically wherever people can connect and share messages, for example, on social media platforms (Hjorth & Hinton, 2019, 19).

The nature of social media has changed significantly over the years, even though it still focuses on user-generated content. Social media is significantly modified by algorithms that seek to identify and predict user actions and behavior, and to guide appropriate information and content to users based on this information. For example, Facebook and Google largely operate by the use of algorithms. The purpose of the algorithms is to profile users and filter the content that is best suited for them while excluding some content from them (Silvola, 2016). Algorithms are explored in more detail in Section 4.

Social media channels also create filter bubbles that isolate like-minded users (Seppänen & Väliverronen, 2013, 212). Indeed, social media filter bubbles are precisely based on the fact that algorithms produce content that is relevant to the user's interests. And since social media has also become a major marketplace, one of the purposes of these filter bubbles is to entice users to spend as much time on the page as possible, because of the advertising revenues (Elo, 2018). It is also important to note that when using social media channels, users provide a lot of information about themselves to social media companies, which then resell this information to advertisers.

Social media channels have also raised concerns when compared to traditional media. On Facebook or Twitter, for example, content can be relayed among users without any third-party filtering or checking the facts. Publications of individual publishers can also reach as many readers as the most popular newspapers, such as the New York Times. (Allcott & Gentzkow, 2017, 211).

In order to better understand how social media has changed the public sphere, we will look at the concepts of the public sphere, algorithms, targeted marketing, social curation and fake news. To get a better picture of the impact of social media on the public sphere, we use the 2016 US presidential election as a case study. Next, we will take a brief look at why the US 2016 election was called a social media election.

## 2.1. Why did social media matter in the 2016 election?

The Internet, social media and mobile apps have become a significant part of people's daily lives. In the US, for example, up to 87% of adults and almost 100% of students use social media and other digital communication technologies daily. In addition, the time spent on social media is also significantly increasing. (Stephen, 2015, 3.)

The interactive nature of social media has also changed the way political campaigns are arranged and organized. Various social media channels such as Twitter and Facebook offer political candidates an excellent way to communicate strategically with potential voters and target audiences. Social media has become the new norm in the US presidential election after President Barack Obama succeeded with the help of social media in the 2008 and 2012 elections (Abdulsamad et. al. 2019, 1).

From the perspective of political campaigning social media channels, such as Twitter, offer a more empowered form of communication. This is particularly seen in the ability of social media to engage and reach a much wider audience than traditional media is able to. Secondly, publications can be easily forwarded. For example, on Twitter, this can be done by retweeting a post. Additionally, campaign staffers can also take advantage of social media to bring more visibility to the campaign and the candidate (Ross & Rivers, 2018, 1-2).

Social media also offers candidates a way to bypass news media gatekeepers and communicate directly with voters (Abdulsamad et. al. 2019, 2). During the 2016 US presidential election Donald Trump and Hillary Clinton showed how social media can be used as a news source for the public.

In that election, social media, especially Twitter, proved to be an important channel for both candidates. However, Trump's use of Twitter was particularly on the agenda of the debate as it was highly unusual in a political context due to the fact that tweets came directly from him. This unconventional approach has continued after his election and has sparked widespread debate about the motives and expediency of his tweets (Ross & Rivers, 2018, 2).

One of the central themes of Trump's tweets during the election was his attacks and derogatory comments about the mainstream media. He used the terms "fake news" and "fake media" to make the public question the reliability of media reporters, especially those who criticized him. At the same time, Trump emphasized his position as the only reliable and truthful source of information. During the election, Trump's use of social media was also analyzed as a source of disinformation and fake news, as his own agenda and goals were best achieved by spreading false information (Ross & Rivers, 2018, 2).

In the 2016 election, the spread of news via social media, and in particular the circulation of false stories, i.e. fake news received special attention. Recent surveys show that 62 percent of US adults receive their news via social media and that in the 2016 election, the most popular fake news were distributed much more on social media than the most popular mainstream news stories. In addition, studies show that many people detect and report fake news but still believe them. These recent findings have led many to question whether Donald Trump would have won the 2016 presidential election if the distribution of fake news on social media had not been so widespread and significant (Allcott, 2017, 212).

Next, we will take a look at the public sphere, and its new forms in the era of social media. Then we will take a closer look at targeted marketing, algorithms and the social curation and popularity of news media outlets on social media, in the light of the 2016 US presidential election.

## 3. Social Media and the Public Sphere – C. Korhonen

In the last decade, after social media gained popularity, there has been a discussion about its ability to revitalize the concept of the Habermasian public (Dahlgren & Olsson, 2007). The traditional Habermasian concept of the national public sphere created by the mass media is said to have evolved into a multi-layered network of social networks that are more important in engaging and activating citizenship. This section discusses how the Internet and, especially social media, have changed Habermas' concept of the public sphere.

### 3.1. A Brief Introduction to Habermas' Public Sphere

According to German philosopher and social theorist Jürgen Habermas, who outlined the historical stages of publicity, public life and democracy require publicity. In his book, *The Structural Transformation of the Public Sphere* (1989), Habermas describes how publicity as its own social sphere was formed at the turn of the 17th and 18th centuries in England. This happened as part of the rationalization process that gave birth to the whole of civil or bourgeoisie society. In feudal society, publicity did not form its own independent domain but existed as so-called representative publicity. The monarch represented its own power over the people, just as the church represented divine authority.

Starting in the 17th century, European society began to change from feudalism to capitalism or bourgeois society, with a bourgeoisie appearing among the monarchs and noblemen (Habermas, 1989, 22-23). In Habermas' analysis, the idea of publicity flourished as a rising bourgeoisie in Europe in the 18th century began to oppose the absolute right of monarchs and nobles by invoking public debate and the power of the best argument (Seppänen & Väliverronen, 2012, 72). The new bourgeoisie consisted of individuals who emerged from the private sphere of their family and their economy to the political public sphere by engaging in public debate. Publicity forums include cafes, salons, newspapers, and magazines. (Habermas, 1989, 28-30).

However, according to Habermas, the critical form of publicity began to degenerate and took on characteristics of feudal publicity in the late 19th century as leading European states evolved into mass societies with public discourse channeled into consumer capitalism. Magazines engaged in one-way communication, and the proper conversation was replaced by mass entertainment. According to Habermas, new rulers began to dominate their subordinates with propaganda, and the debating public had become a media-consuming audience.

### 3.2. Social Media as a Revival of the Public Sphere?

During the era of digitalization, various social networking sites and platforms have become to form a central part of the public sphere. The Internet and social media have been said to revive Habermas' democratic publicity on a global scale. This idea is based on the view that social media emphasizes interaction, openness, and community (Seppänen & Väliverronen, 2012, 75).

The rise of social media has not only created a new public sphere but also shifted the arena in which political debate happens. Habermas places publicity between the private sphere and the public authorities, an area of civil society where citizens can form opinions and engage in political activities (Habermas, 1962, 30). Open publicity provides an opportunity for the political organization of society. Social media makes this easier than ever. According to Iosifidis (2011, 5) "in theory, this open, free and decentralized space could create the conditions for ideal speech and enhance the ability to voice one's opinion and organize action (the very notion of democracy)."

The Internet and social media can be seen to expand Habermas' view of civil society, because in social media "the formation of public opinion takes place in a transnational context that crosses national boundaries" (Iosifidis, 2011, 5). This means, for example, that people around the world can follow and participate in the country's internal political debate. The presidential elections in the US were also monitored closely in Finland. Speculation about the election and its twists and turns were discussed both in traditional and social media.

Social media has the power to influence public opinion and what is being discussed outside of the Internet. In the 2016 presidential election, Twitter played a significant role in Donald Trump's victory. Twitter was at the center of Trump's campaign and through it, Trump got a lot of free media visibility. Trump's constant posting on Twitter and his provocative comments attracted a lot of media attention and increased conversation around him, which added to his visibility even more (Francia, 2018). According to Francia (2018) "Trump was more often the topic of personal conversations than Clinton."

However, open participation in social media can lead to chaos because there is no structured conversation (Iosifidis, 2011, 6). The social media debate may give rise to more anarchy than democracy (Iosifidis, 2011, 6) because in social media the argumentation is polarized (Dahlberg, 2007). Social media algorithms limit the content that people are exposed to and often displays content that favors an individual's social world. This creates social bubbles which lead to a homogenous atmosphere. Social media does not necessarily promote rational, critical and political debate between citizens. (Dahlberg 2007; Allcott & Gentzkow, 2017.)

In addition, a lot of propaganda and fake news also circulates on social media. During and after the 2016 election, Trump deliberately attacked journalists and scientists by claiming that their studies and news were fake. Francia (2018) cites research in his article that suggests that Trump's tweets made his followers less likely to believe any negative or criticizing news coverage about Trump. Making and spreading false claims could prevent educated political decision-making and reduce voter's selection based on genuine information. (Allcott & Gentzkow, 2017.)

Although it is generally said that the Internet offers everyone the same opportunity to participate in public debate, this is not actually the case. Like Habermas' publicity, the Internet and social media are not open to everyone. In Habermas' view, only men in the upper class were allowed to take part in the debate during its Enlightenment era incarnation. This exercise of debate and discussion excluded women and members of the lower classes. Nowadays the challenge is the limited access to the Internet and its platforms. Not everyone around the world has the same access to the Internet because of censorship or network performance (Iosifidis, 2011, 6-7).

According to Habermas' theory (Habermas 1989), we can say that social media has the potential to serve as a public sphere. As stated earlier, Habermas placed publicity between private and public power, existing in its own public and social sphere. It has also been stated that social media has blurred these boundaries. We can often see in social media how private and public life, leisure, and work merge together (Fuch, 2014), and this does not fully fit with Habermas' idea of a clear distinction between private and public spheres characteristic of the eighteenth century. Also, according to Fuch (2014), social media does not currently form the public sphere because it is not free from political or economic power.

Economic interests and advertising shape social media in many ways (Seppänen & Väliverronen, 2012, 75). The most popular social media sites sell their users' information to advertisers and those in power who use the information to, for example, deliver political campaign messages. Google, Facebook, Twitter, and YouTube all sell their users' public and private information to advertisers (Fuch, 2014). Social media sites make a profit from user-generated content. Also, through social media and its algorithms users can be monitored more easily, taking surveillance to a new level (Fuch, 2014).

In the next section, we will discuss social media algorithms in-depth and explain how they have affected the public sphere.

## Social Media Algorithms – L. Trémouille

In the era of the Internet and social media, algorithms have had a significant effect on the content we see and consume. In this section, we will explain how social media algorithms work and how they shape the public sphere.

### 4.1. What are algorithms?

> We now exist in these curated environments, where we never see anything outside our own bubble … and we don't realize how curated they are.
> —Emily Taylor, a chief executive of Oxford Information Labs and editor of the *Journal of Cyber Policy* (Hern, 2017).

Algorithms are often described as a set of instructions or as a recipe. Basically, a programmer tells the algorithm how to operate if a specific set of conditions is met. This process is described as "IFTTT", which stands for "if this then that" (Hildebrandt, 2016, 1). With social media algorithms, the process could be for example, "if a user likes this, comments on that, and opens this article, then show more of this kind of content".

Lately, as artificial intelligence has developed significantly, machine learning has gained a lot of popularity among programmers. Machine learning is a sub-discipline of AI, in which AI is programmed to reconfigure its own behavior to improve its performance based on the data it processes (Hildenbrandt, 2016, 1). With this development AIs move from following strictly given IFTTT-rules to develop their own algorithms autonomously. Therefore, the programmer of an AI that is based on machine learning cannot be in full control of how the algorithm will eventually work. These two types of algorithms are also called automatic (IFTTT) and autonomous (machine learning) (Boulanin, 2019, 20).

### 4.2. How do social media algorithms work?

Social media has become an inseparable part of everyday life for many modern-day people. It has changed the way we communicate and form communities by providing access to ideas online without limitations of time or space (Delacruz, 2009). As stated in Section 2, 62 percent of US adults receive their news via social media. However, what seems to be unclear to many is how the algorithms work and, therefore, how the content received through social media is curated to each user, based on the algorithmic analysis of that user (Association for Computing Machinery US Public Policy Council (USACM), 2017).

As social media platforms have gained popularity among users, the content and information on the platforms have overflowed. To overcome this issue, social media companies use recommender systems to suggest useful content for each user. Recommender systems analyze users' behavior and compare it to the data of other users. They create a profile of the user and then recommend content based on what neighbor users (users with similar profiles) have shared or liked (Anandhan et al., 2018). There are several different types of recommender systems such as filtering, collaborative filtering, and hybrid-based filtering, but what all of these have in common, and what is important in the light of this article, is that all of these systems somehow limit the content and information the user receives when using social media services.

Algorithms function between the user and the information and they can be perceived as both the interpreter and the gatekeeper. From a positive perspective, algorithms know and understand the user's preferences, sometimes better than the user themselves. From this point of view, algorithms are to some extent reading the user's mind and acting based on that; when a user types "Donald" to Google's search engine, it automatically suggests "Donald Trump" and if that is what the user is searching for, the algorithm can be seen as guessing the user's intention. Although for many this is a positive functionality, it also filters out some of the content, and therefore algorithms can be seen as the new gatekeepers of information (Roth, 2019, 3). This functionality of the algorithms also creates filter bubbles or echo chambers, where the existing ideas and values of the user are strengthened without providing any new or contradicting information.

As stated earlier, everything a user sees on social media is curated by algorithms. This curation is done based on the users' explicit declaration about their preferences, for example, joining a certain group or liking a certain page, but also their implicit behavior by analyzing the user and comparing their profile to other users (Roth, 2019, 4). One major factor that determines what the user sees is money, as most of the social media platforms are funded by displaying ads (Hern, 2017). What is also important to keep in mind when discussing social media algorithms is that the revenue of social media companies is based on the time that the user spends on their websites, viewing the displayed ads.

Social media is often portrayed and viewed as an enabler of communication or as a platform for user-based content. What is often overlooked is that behind every social media service there is a company that is trying to make money. The core business of the companies such as Facebook is to collect data on its users and then use that data to provide advertisers direct passage to reach their target groups (Benkler et al., 2018, 269). As the 2016 presidential election proved to a large audience, this also provides a way for political candidates to reach their possible voters with targeted messages.

### 4.3. How can social media algorithms shape the public sphere?

Many recent studies have shown that the recommendations made by algorithms have a significant influence on what we consume, purchase or believe. They affect the music we listen to on Spotify, movies we watch on Netflix or products we purchase from Amazon, as all of these services give us recommendations (e.g. Hosanagar, 2019).

After the 2016 election, the influence of algorithms has been tested on several occasions. For example, Epstein and Robertson (2015) tested how the order of the results on a Google search results page influences the popularity of Australian political candidates among US citizens who were unfamiliar with the candidates before the experiment. They found that people tended to trust the results that were displayed first more, and that they formed their opinions based on these results.

Facebook researchers have also found that by providing users a way to inform their friends they have voted by adding functionalities such as an “I voted” button, they could potentially urge other users to vote as well. Although this can be seen as a positive development, it also has the potential to favor certain candidates or parties depending on which users receive this “nudge” (Caplan & Boyd, 2018, 1). Gillespie (2010, 347) has described algorithms as “curators of public discourse” as they have such a significant impact on how the public discourse appears to each user.

Algorithms can, therefore, have a major impact on politics and the public sphere. Thus, social media companies and other online platform providers have a great responsibility when developing these algorithms and services. If the algorithms or the data provided to the algorithms is biased, it has the potential to change public life and possibly a society, as we have seen in the US. However, as Caplan and Boyd (2016) point out, biased algorithms are not the only concern when it comes to algorithmic influencing of the public sphere, because algorithms can also be manipulated by those who find ways to manipulate the information flow.

## 5. Targeted Marketing – S. Siirtola

Social media has brought new forms of marketing and advertising into our daily lives. Digital marketing provides new ways to advertise goods and services online and on different social platforms. One of the key factors in digital marketing has been personalized ads that are tailored for each specific user. This means that every user online has ads showing up specifically targeted for them. This kind of targeting is based on information and data gathered from users. The data includes information on which sites you visit, what links you click on and what kinds of purchases you make online, among other things.

The basic idea of targeted marketing is to identify groups that an organization could target (Lancaster, 1988, 79). This is based on a practice wherein marketing companies buy consumer information and data that is collected on sites such as Facebook. People enter their personal information on their Facebook profile which then Facebook harnesses and sells to advertisers. The problem here is that the consumers do not always know how their information is being used, which is a source of privacy concerns.

Targeted marketing is seen as more effective than so-called offline marketing because it allows companies to focus on their target groups. In the book *Digital Marketing Playbook*, Ian Dodson talks about using digital marketing to engage in a dialog with the consumers and thus help them communicate with the company. Dodson mentions a term called DDA which stands for digital display advertising (Dodson, 2016, 91). It is a form of digital marketing that displays ads online as a means of communicating relevant commercial messages to a specific audience based on their profile (Dodson, 2016, 91).

Dodson talks about the trail users leave online while using social media platforms and generally browsing online (Dodson, 2016, 99). The users' clicks give you information on who they are, where they live and what they are interested in. He uses an example of a woman called Debbie to show what kind of information can be collected. Debbie can be identified according to her age, marital status, hobbies, and income. It becomes easier to target ads for the person when you know all this basic information about her, her interests and her consumer behavior (Dodson, 2016, 99).

There is also another issue with targeted content: the previously mentioned social bubbles. Creating targeted content restricts the content users view and receive. When people search for something online, everyone gets different kinds of results and this affects the content we see online. The same happens for example on Facebook: we mostly see the posts of the people with whom we are most connected. This can also shape our ideologies or make them stronger.

In the article "Social Media and Fake News in the 2016 Election," Allcott and Gentzkow mention a study conducted in 2015 which shows that Facebook friend networks are ideologically segregated (Allcott & Gentzkow, 2017, 221). Because users are mainly seeing the posts of the people with whom they are the closest with and share the same ideologies with, they get fewer posts about opposing ideologies (Allcott & Gentzkow, 2017, 221). People are thus more likely to read and share news articles that are aligned with their own ideologies. According to the article, this kind of segregation suggests that "people who get news from Facebook (or other social media) are less likely to receive evidence about the true state of the world that would counter an ideologically aligned but false story" (Allcott & Gentzkow, 2017, 221). This was also an important factor in the presidential election in 2016.

### 5.1. Targeted Marketing in the Election of 2016 - Cambridge Analytica

In the US presidential election of 2016, social media played a big role in the campaigning of the candidates. As mentioned earlier, the key element in Trump's campaign seemed to be Twitter and other social media channels. Trump managed to target his campaign to the right groups even though his methods were questionable. This included talking about Clinton offensively on public platforms, trying to make her look bad.

One thing, in particular, raised a lot of conversation in the media and among the public. It was the way Trump's campaign had collected data on millions of people on Facebook which was then used in his campaign. The data the Trump team collected allowed them to tailor digital ads and online fundraising efforts to specific users (Prokop, 2018).

Trump used a company called Cambridge Analytica to collect data on possible voters. The CEO of the company commented on the case saying, "we did all the research, all the data, all the analytics, all the targeting, we ran all the digital campaign, the television campaign, and our data informed all the strategy" (Prokop, 2018). An academic at Cambridge University, called Aleksandr Kogan, created an app for Facebook to collect the data. The app was called “thisismydigitallife” and it was a personality quiz which users could take on Facebook. To take the quiz, users had to consent to give the app access to all of their and their friends’ Facebook profiles. In the end, more than 270 000 people took the quiz in the app which lead to over 50 million profiles to be collected and over 30 million people were identified on the electoral rolls (Prokop, 2018).

The company used the data to build an algorithm that could analyze individual Facebook profiles and determine personality traits linked to voting behavior. The algorithm and database formed a powerful political tool together. This allowed the campaign to identify possible swing voters and craft messages that were more likely to resonate. This managed to create profiles that were identifiable and tied to electoral registers, across 11 states, with the possibility to expand even further. The 50 million profiles gathered represented around a third of active North American Facebook users, and nearly a quarter of potential US voters (Cadwalladr & Graham, 2018).

Facebook as a company played a crucial role in the campaign, as well. Facebook provided an interface that allowed campaigns to target specific voters, geographic regions, or demographics, as well as to send ads to the users based on their personal data (Benkler, 2018).

The main goal of the data used in the campaign was to predict and influence choices at the ballot box by targeting potential voters with personalized political advertisements. This raises serious questions about Facebook’s role in targeting voters in the US presidential election (Cadwallader & Graham, 2018). According to Benkler, the fundamental problem is that Facebook’s core business is to collect highly refined data about its users and convert that data into “microtargeted manipulations” such as advertisements and newsfeed adjustments (Benkler, 2018). The purpose of this is to get its users to want, believe, or do things. Benkler says that “that same platform-based, microtargeted manipulation used on voters threatens to undermine the very possibility of a democratic polity” (Benkler, 2018, 269). Cambridge Analytica took advantage of the intentional design of the basic business of Facebook and turned it into a political tool. According to Benkler, when this is applied to political communication, it presents a long-term threat to democracy (Benkler, 2018).

## 6. Social Curation: Popularity of News Media Outlets on Social Media – M. Liukkonen

When talking about the 2016 election and the powers of social media, we need to understand how Americans receive their political information and how these ways have been changing before and during the election.

When observing the election, attention needs to be paid to the architecture of discontent and how “bubbles” are formed in the media. The digitalization of news media cannot be ignored, either. Even though “fake news” is not a new term, it gained widespread publicity during Trump’s and Clinton’s presidential campaigns. As cited before, it has been wondered whether if the distribution of fake news on social media had not been so widespread, Donald Trump would have won the 2016 presidential election (Allcott & Gentzkow, 2017, 212)?

### 6.1. The architecture of our discontent

In America, people get their information about politics from a very diverse set of sources. Even though more and more Americans use social media as their primary source of news, a large portion still relies on broadcast television and cable news. That is why in order to understand media, especially in the context of American politics, we need to understand the entire ecosystem (Benkler, Faris & Roberts, 2018, 45).

It has been observed that there is no left-right division in American media, but rather a division between the right and the rest: the American media ecosystem is divided into two distinct and structurally different ecosystems. Besides the right-wing media, the rest of the spectrum includes outlets from the left and center-right publications. This ecosystem is often seen to adhere to professional standards of journalism (Benkler, Faris & Roberts, 2018, 73-74).

Because of the strict division, the way people receive their media content is asymmetric—even when we talk about traditional mass media such as television and cable news. For example, according to surveys, respondents who identified as "consistently conservative" politically, also reported that their most trusted source of news was Fox News, Sean Hannity, and Rush Limbaugh. "Consistently liberal" people responded that their most trusted media were NPR, PBS, and the BBC. Even though these patterns were observed about television and radio sources, the same pattern was also congruent online (Benkler, Faris & Roberts, 2018, 73).

Because of this asymmetry, social bubbles are strong, and people are receiving messages only from similarly-minded people. Adding social media to these already visible patterns in the traditional mass-media, people can curate what they read and see. This leads to a situation where people seek out evidence that is in alignment with their preconceptions, interact only with like-minded people, and avoid information that does not fit with what they believe and like to hear confirmed (Benkler, Faris & Roberts, 2018, 289-290). The problems presented by this asymmetry increase when the political polarization rises and there are more and more negative feelings on each side of the political spectrum towards each other (Allcott & Gentzkow, 2017, 214). Therefore, like-minded messages are exchanged even more widely without regard for the facts.

### 6.2. The digital transformation of news media and the rise of disinformation and fake news

Besides traditional media and the asymmetries therein, digitalization of the news media has also changed the way Americans received political information during the 2016 election. Even though a large portion of Americans used television and radio as their main source of news, social media use has risen sharply. In 2016, Facebook had

1.8 billion active users per month and Twitter had almost 400 million active users each month (Allcott & Gentzkow, 2017, 214).

In the digital age anyone, anywhere, can produce content. Agendas are diverse, signals are mixed, and messages manipulated. Barriers of entering the media industry have almost disappeared because it is easy to create media outlets (for example websites), as well as to monetize content through advertising (Allcott & Gentzkow, 2017, 214-215). The increasing amount of different online news providers has shaken the media ecosystem and raises quality concerns about news channels.

Technological changes such as algorithms and changed media logics have had a strong impact on the news media industry. Instead of direct access to newspapers, people tend to have algorithm-driven access to unbundled articles. Therefore, journalists and newspaper editors lose control of the curation of the content (Martens, Aguiar, Gomez-Herrera & Mueller-Langer, 2018, 12-16). Wider access can be seen as a good aspect, but it also might affect the perceived quality of news. Because of the shape of the online news production and very short production cycles, there is no time for constant fact-checking and unchecked news is more common even in high-quality newspapers (Martens, et al., 2018, 18).

This so-called digital transformation can be seen as a reason why especially social media platforms are conducive to fake news. First, entering the new media market and producing content is cheap. Also, it is difficult to judge the messages' veracity because of the format: thin slices of information viewed on phones. Thirdly, social networks are ideologically segregated, and people are more likely to read and share the news that is aligned with their ideology (Allcott & Gentzkow, 2017, 221).

The digitalization and new production cycles can be seen as one of the aspects that affect the quality of news, but changed media logic cannot be blamed alone. Social media has created a platform for fake news, for sure. However, social media do not create fake news. The motivations behind the creation of fake news can be divided into two categories: pecuniary and ideological. Content that is popular, or even viral, on social media can draw advertising revenue when linking to the original site. The other side is ideological and can be seen best in politics: some fake news providers seek to advance their candidates (Allcott & Gentzkow, 2017, 217.) The Internet and social media have destabilized traditional institutions, including television and print news outlets. Hence, marginalized voices and messages are allowed to reach out directly to audiences. The downside is, that the same phenomena allow actors to distribute propaganda and fake news (Benkler, Faris & Roberts, 2018, 289).

Digital transformation and political polarization created a platform for the 2016 election in which Trump was able to use social media to achieve his own goals by spreading false information. During the election, Trump's use of social media was also analyzed as a source of disinformation and fake news, as his agenda and goals were best achieved by spreading false information (Ross & Rivers, 2018, 2).

## 7. Conclusion

Social media’s ability to provide opportunities for content production without any third-party filtering or verification of facts has sparked concerns. For example, publications of individual publishers can gain more visibility than posts by the most popular traditional media outlets. This is a very good example of why and how social media has changed the public sphere, and this was clearly visible in the 2016 US presidential election.

In that election, both candidates, Trump and Clinton, took advantage of social media in their campaigns, as they were aware of its ability to bypass news media gatekeepers and communicate directly with voters.

Donald Trump's Twitter behavior during and after the election sparked debate and it has been analyzed as a source of disinformation and fake news, since Trump's tweets focused on his attacks and his derogatory comments on the mainstream media and his opponents.

Many have questioned whether Donald Trump would have won the US 2016 presidential election if he had not been so active on social media during the election. For these reasons, it is important to note that social media may have had a major impact on the results of the US presidential election and the contemporary public sphere more generally.

As shown, the public sphere has changed and expanded much since social media gained popularity. Through social media, almost everyone has the opportunity to engage in public and social debate, regardless of the country's borders. This idea broadens the views of Habermas’ public sphere because social media can increase people's political participation.

During the 2016 American presidential election, the role of social media in the arena of political debate became central. Facebook, YouTube, and Twitter were a key part of the public debate. However, deciding whether social media represents a revival of Habermas’ public sphere is not easy. Social media is shaped by a wide variety of economic and political agendas. Not everyone around the world has the same access to the Internet and social media services. In addition, social media exercises a lot of control and censorship over its users. It is used not only for safety but also for marketing purposes.

Targeted marketing has become an effective tool for finding the right audiences online. It can be used to create personalized ads and promote sales targeted to a certain group. It can also make it easier for people to browse online and find the things that they are the most interested in. However, when applied to public debate, targeted marketing can reduce rational political debate and decision-making as users mostly encounter like-minded views.

Additionally, targeted marketing is mostly based on collecting information of online users, and people cannot always be sure what their data is used for. As we can see in the Cambridge Analytica case, the data collected can be misused for political purposes, and this makes the use of targeted marketing somewhat questionable.

The mainstream American media ecosystem is divided into two polarised ecosystems: right-wing media represented by Fox News, and left-wing media exemplified by MSNBC and sometimes interpreted to include CNN. Because of this division, the way people receive their media content is asymmetrical, meaning that a big portion of the population receives their political information from like-minded media channels. Of course, media exposure and viewing habits vary by generation, so more research is needed to address various segments of the public.

In addition, the digitalization of news media has changed the ways in which actors can deliver news content, giving more space to marginalized voices. Almost anyone can produce content anywhere at no cost, but without fact-checking. This change has also made it easier to produce disinformation and fake news.

## References

Allcott, H. & Gentzkow, M. (2017). Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives, 31*(2), pp. 211-236. doi:10.1257/jep.31.2.211

Anandhan, A., Shuib, L., Ismail, M. A. & Mujtaba, G. (2018). Social Media Recommender Systems: Review and Open Research Issues. *IEEE Access, 6*, pp. 15608-15628. doi:10.1109/ACCESS.2018.2810062

Benkler, Y. (2018). *Network propaganda: Manipulation, disinformation, and radicalization in American politics*. New York, NY: Oxford University Press.

Boulanin, V. (2019). *The impact of artificial intelligence on strategic stability and nuclear risk*. Solna, Sweden: Stockholm International peace research institute.

Cadwalladr, C., & Graham-Harrison, E. 2018. "The Guardian, Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach 2018". The Guardian. Published 17th of Mar 2018. [Viewed on 4th Nov 2019]

Caplan, R. & Boyd, D. (2016). *Who Controls the Public Sphere in an Era of Algorithms*?. New York, NY: Data & Society.

Dahlberg, L. (2007). Rethinking the fragmentation of the cyberpublic: From consensus to contestation. *New Media & Society, 9*(5), pp. 827-847. doi:10.1177/1461444807081228

Dahlgren, P., Olsson, T. & Butsch, R. (2016). From public sphere to civic culture: Young citizens' Internet use. *Media And Public Spheres,* pp. 198-209. doi:10.1057/9780230206359

Delacruz, E. M. (2009). From Bricks and Mortar to the Public Sphere in Cyberspace: Creating a Culture of Caring on the Digital Global Commons. *International Journal of Education & the Arts, 10*(5).

Dodson, I. (2016). *The art of digital marketing: the definitive guide to creating strategic, targeted, and measurable online campaigns*. John Wiley & Sons.

Elo, E. (2018) ”Vaalit, joiden takia voi tulla tapetuksi − Yhdysvaltain vaalijärjestelmä on villi länsi, ja hakkeri Harri Hursti on sen keskellä”. *Kauppalehti*. 7.11.2018. Viewed: 18.10.2019.

Epstein, R. & Robertson, R. E. (2015). The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. *Proceedings of the National Academy of Sciences of the United States of America, 112*(33), p. E4512. doi:10.1073/pnas.1419828112

Francia, P. L. (2018). Free Media and Twitter in the 2016 Presidential Election: The Unconventional Campaign of Donald Trump. *Social Science Computer Review, 36*(4), pp. 440-455. doi:10.1177/0894439317730302

Fuchs, C. (2014). Social Media and the Public Sphere. *tripleC: Communication, Capitalism & Critique, 12*(1), pp. 57-101. doi:10.31269/triplec.v12i1.552

Garfinkel, S., Matthews, J., Shapiro, S. & Smith, J. (2017). Toward algorithmic transparency and accountability. *Communications of the ACM, 60*(9), p. 5. doi:10.1145/3125780 [viewed on 3th November 2019]

Gillespie, T. (2010). The politics of 'platforms'. *New Media & Society, 12*(3), pp. 347-364. doi:10.1177/1461444809342738

Habermas, J. (1989). *The structural transformation of the public sphere: An inquiry into a category of bourgeois society*. Cambridge: Polity Press.

Hern, A. (2017). How social media filter bubbles and algorithms influence the election. *The Guardian*, *22*, 2017. [Viewed on 3th Nov 2019]

Hildebrandt, M. (2016). The new imbroglio. living with machine algorithms. *The Art of Ethics in the Information Society*, 55-60.

Hinton, S. & Hjorth, L. 2013. *Understanding Social Media.* London: SAGE Publications Ltd. DOI: 10.4135/9781446270189.

Hosanagar, K. (2019). *A Human's Guide to Machine Intelligence: How Algorithms are Shaping Our Lives and how We Can Stay in Control*. Viking.

Howard, P. N., Woolley, S., & Calo, R. (2018). Algorithms, bots, and political communication in the US 2016 election: The challenge of automated political communication for election law and administration. *Journal of information technology & politics*, *15*(2), 81-93.

Iosifidis, P. (2011). THE PUBLIC SPHERE, SOCIAL NETWORKS AND PUBLIC SERVICE MEDIA. *Information, Communication & Society, 14*(5), pp. 619-637. doi:10.1080/1369118X.2010.514356

Lancaster G. & Massingham, L. (1988). *Essentials of Marketing.* New York, NY: McGraw-Hill Publishing Co.

Martens, B., Aguiar, L. & Muller, F. (2018). The digital transformation of news media and the rise of disinformation and fake news. European Commission JRC Technical Report.

Prokop, A. 2018. "Cambridge Analytica shutting down: the firm's many scandals, explained". Cambridge Analytica shutting down: the firm's many scandals, explained 2018. *Vox.xom.* Published 2nd of May, 2018. [Viewed on 4th Nov 2019]

Ross, A. S. & Rivers, D. J. (2018). Discursive Deflection: Accusation of "Fake News" and the Spread of Mis- and Disinformation in the Tweets of President Trump. *Social Media + Society, 4*(2), . doi:10.1177/2056305118776010

Roth, C. (2019). Algorithmic Distortion of Informational Landscapes. *Intellectica, 70*(1), pp. 97-118.

Sahly, A., Shao, C. & Kwon, K. H. (2019). Social Media for Political Campaigns: An Examination of Trump's and Clinton's Frame Building and Its Effect on Audience Engagement. *Social Media + Society, 5*(2). doi:10.1177/2056305119855141

Seppänen, J. & Väliverronen, E. (2012). *Mediayhteiskunta*. Tampere: Vastapaino.

Silvola, S. 2016. ”Vaikuttiko Facebookin algoritmi Yhdysvaltain vaalitulokseen?”. *YLE* 9.11.2016. [Viewed on 18th Oct 2019]

Stephen, A. T. (2016). The role of digital and social media marketing in consumer behavior. *Current Opinion in Psychology, 10*(C), pp. 17-21. doi:10.1016/j.copsyc.2015.10.016

# 2.3 The Digital Public Sphere and the Presidential Campaign of Ukrainian President Zelensky

Mio Tamakoshi, Inka Reinola, Sakari Nuuttila, Laura Kuokkanen,
Vanessa Garcia Torres
Faculty of Social Sciences, University of Helsinki

## Abstract

The paper will examine the digital public sphere applied to the case study of the recent Ukrainian presidential elections of April 2019. In the first section, it will reflect on what Jürgen Habermas considers as the public sphere in order to establish the basis of the investigation. The second section will define the current role of the public sphere in the digital age. In the third section, we will move to the case, focusing on the campaign of the elected president Volodymyr Zelensky, and describe the fragmentation of the public sphere in relation to the messages addressed to different audiences during the campaign. In the fourth section, we will further reflect on the manifestation of the digital public sphere in Zelensky’s entirely digital campaign, as well as in the seeping of fiction into reality, as Zelensky became renowned for playing the lead in the tv series "Servant of the People”, about a normal man who inadvertently becomes president of Ukraine. Finally, we will examine how the digital public sphere influenced the Ukrainian presidential election and the effects it has had.

*Keywords*:  Public sphere, democracy, deliberative democracy, elections, campaign, digital public sphere, Ukraine, Eastern Europe, Zelensky, social media

This essay will analyse the 2019 Ukrainian presidential election applying the concept of the public sphere. With the impact of the digital era the concept of the public sphere should be analysed in a new context. The 2019 Ukrainian presidential election showed how the new public sphere in the digital era functions and what kind of implications it has on deliberative democracy.

First, the essay will look at the extension of the concept of the public sphere in Jürgen Habermas' theoretical works. This section attempts to reveal the extent to which the concept of the public sphere is intended to recommend aspirational norms, and thus assess how applicable it is to the digital age and to the concrete case of Ukraine.

In the next section the essay investigates the concept of the digital public sphere and the special features of the Ukrainian public sphere in the digital era. This section tries to answer the questions of what the concept of the digital public sphere signifies and why it is important, as well as what the features of Ukrainian digital public sphere are.

In the third section, the essay will delve into the Ukrainian election and the reasons for Volodymyr Zelensky's victory through the theoretical framework of the fragmentation of the public sphere. In addition to Habermas' theory, this essay will present different views on this fragmentation by relevant scholars. This section will also investigate how Zelensky possibly used a divided public society to his advantage, whether the fragmentation of the public sphere explains the events in Ukraine sufficiently, and whether Zelensky added to the fragmentation.

Fourth, the essay will concentrate specifically on the digital aspect of Zelensky's presidential campaign, which was largely based on his fictional television persona and was conducted in an entirely virtual manner, and therefore works as a prime example of the manifestation of the digital public sphere in contemporary political reality. Zelensky's campaign, which fostered emancipatory potentials of the digital public sphere, but less so deliberative ones, is also an illustrative example of both the participatory potentials and shortcomings of the digital media environment as a contemporary public sphere.

Finally, the essay describes the influence of the digital public sphere in the Ukrainian presidential election and its effects during the campaign. It analyses the digital strategies used by Servant of the People Party and current President Volodymyr Zelensky. In addition, it addresses the impact of social networks platforms to express ideas and opinions to electors and citizens. A clear comparison between Zelensky's campaign and other similar cases, like American presidential elections held in 2008 and 2016, are also presented.

## 1. The extension of the concept of the public sphere by Habermas

This section will investigate how the concept has been extended in Jürgen Habermas' own theoretical works. This question attempts to serve as the basis for the following sections by showing the extent to which the concept of the public sphere is intended to be normative and thus how applicable it is to the digital age and to the concrete case of Ukraine.

Nearly half a century has passed since Habermas' concept of the public sphere (*Öffentlichkeit*) was introduced in *Structural Transformation of the Public Sphere* (1962, 1989), and it has now become a common reference across different disciplines. While the German term *Öffentlichkeit* is also used in ordinary language, in English the translated concept "public sphere" is used in politics, history, and sociology to carry a derivative meaning. Habermas himself also modified the usage of the word following this expansive trend. In these circumstances, researchers using the concept of the public sphere are free to use it as a general term, often without a particular reference to Habermas (Calhoun, 1992).

In his *Structural Transformation*, Habermas gives no clear definition of the concept of the public sphere. Rather, he focuses on the etymological explanation as follows. *Öffentlichkeit* is a noun made from the older adjective modeled after French *publicité* and English publicity in the 18th century (Habermas, 1989, p. 2). The concept of "public life" *(bios politikos*) originated in ancient Greece, was taken into Roman law and succeeded as a definition to distinguish the public from the private and to define the region of *res publica*. Habermas cites these conceptions from Hannah Arendt's *The Human Condition* (Habermas,1989, pp. 19-21). However, there seems to be a conceptual transition in addressing the bourgeois revolution period of Britain, the United States, and France as an era of the establishment of the public sphere, and referring to Arendt.

Analyzing the distinction between the public and private realms in political life in the ancient Greek polis, Arendt (1958) claims the re-establishment of a domain as the public realm. She argues that responsible individuals, as equal participants in the political life of the *polis*, were able to form direct participatory political discussions in which they openly expressed their political opinions and argued on an equal footing. However, according to her, given the increasing numbers of individuals with social status in the 18th and 19th centuries, the realm of the social in the modern world finally reached the point where "it embraces and controls all members of a given community equally and with equal strength" (Arendt, 1958, p. 41). Here the public is only ruled by interest, and the subject of this interest is calculated not as individuals nor as their will and opinions, but as a totaled and quantified homogenic subject under bureaucratic administration. For Arendt, this transformation means degeneration and alienation of mankind, all of whom aspire to solemnly express their opinions in public life (Arendt, 1958, pp. 320-5).

According to Habermas' articulation, in contrast, the public sphere, which was established during the bourgeois revolutions of Britain, the United States, and France, has degenerated since the 19th century along with the progress of the industrial revolution (Habermas, 1989, pp. 73-9). In other words, as the system was divided into politics led by political parties that worked to achieve universal suffrage and women's suffrage, in the process of industrialization, the public sphere became a place for advertisements for consumer goods (merchandise) and cultural products retailed by companies, and a site dominated by mass communication as an institution of political propaganda by parties. This theoretical line is derived from the Frankfurt school, to which Habermas belonged, and was nurtured in the context of the criticism of instrumental reason in the mass society and modern communications (Horkheimer & Adorno, 1947).

Therefore, the bourgeois revolutions in Britain, the United States, and France, which Habermas takes up as the establishment of the public sphere, was for Arendt an era of the rise of the social and an era of the degeneration of the public sphere. For Arendt, the structural transformation of the public sphere had already ended at the time of the emergence of the public sphere in Habermas' sense. Hence it can be thought that this earlier decay of the public sphere diagnosed by Arendt was built into his argument from the beginning as a way of introducing the idea of the civil society and further developing the usage of the term "public sphere".

If one sets aside the distinction above and accepts Habermas' understanding of history, the public sphere established during the bourgeois revolution period subsequently disappeared because it is necessarily a phenomenon of historical non-reproducibility. Taking that into account, the public sphere cannot be discussed any longer in terms of its implementation at the present time, which has undergone structural transformation. Nevertheless, it should be noted that, according to Calhoun, Habermas analyses "the historical category of the public sphere and attempts to draw from it a normative ideal" (Calhoun 1992, p. 39). Since it was proposed in 1962, his concept has equivocated between historical actualities and theoretical ideals. The latter, universal and normative aspects can be detected by positioning the modified concept of the public sphere in his theoretical works after his influential text was published.

From *Structural Transformation of the Public Sphere*, through *Theory of Communicative Action* volumes I and II (1984, 1987) to *Between Facts and Norms* (1992), Habermas' understanding of the public sphere has shifted from the view laid out in *Structural Transformation* based on the negative evaluation of the mass media influenced by the Frankfurt School's criticism of instrumental reason. First of all, in *The Theory of Communicative Action*, following Durkheim, Mead and Weber's examinations, the media is portrayed in a negative way as a cultural industry, but with ambivalent expectations for it as a medium of citizen discussion (Habermas 1984, pp. 389-90). The preface to the second edition of *Structural Transformation* expands the concept by referring to Adam Ferguson's idea of civil society, considering that the public sphere of the first edition was the bourgeois public sphere which was realized by the bourgeois revolution (Habermas, 1991, pp. xviii-xix). Furthermore, in *Between Facts and Norms* (1992) Habermas lays out expectations for the public sphere as a forum for discussion on common good and social justice, and the form was taken to match the examples observed in modern society:

> Moreover, the public sphere is differentiated into levels according to the density of communication, organizational complexity, and range—from the episodic publics found in taverns, coffee houses, or on the streets; through the occasional or "arranged" publics of particular presentations and events, such as theater performances, rock concerts, party assemblies, or church congress; up to the abstract public sphere of isolated readers, listeners, and viewers scattered across large geographic areas, or even around the globe, and brought together only through the mass media. (Habermas, 1996, p. 374)

At this theoretical point, the concept of the public sphere has explicitly become applicable to an analysis of a wide range of arenas of public debate that discuss public issues related to people's lives. Hence, as Habermas' own theoretical development suggests, the historical particularity seen in the original conceptualization can be set aside and the universality and normativity of the concept can be utilized to apply to analysis irrespective of time and space.

## 2. The digital public sphere in Ukraine

After exploring Habermas' theory on the public sphere, we next shift to the idea of the public sphere in the digital age. According to Hacker and van Dick (2000), digital democracy is defined as "a collection of attempts to practice democracy without the limits of time, space and other physical conditions, using…[new] ICT [Information & Communications Technology]…as an addition, not replacement, for traditional analogue practices" (cited in Royston & Creeber, 2009, p. 139). The Internet is usually considered as an open platform and hyper-interactive medium; it can be argued that the Internet is generally a relatively open and accessible public sphere (Royston & Creeber, 2009, p. 141). On the other hand, some countries have heavily controlled the Internet and it could be argued that Internet freedom or democracy of the digital world is not actualized because of the restrictions on its openness. However, the Internet has been used as an emancipator even by controlling political regimes, since the growth of access and the use of ICT in the Middle Eastern and African region "provided resources for protesters and governments to influence political and social events", which played a role in the Arab spring in the early 2010s (Wilson & Corey, 2012, p. 343).

What are the characteristics of a digital public sphere, when we look beyond the traditional public sphere? What characterizes the Internet is its interactivity, which could be argued as going beyond the traditional public sphere. The "Internet is characterized as a lateral, interactive and discursive model of communication" in opposition to for example radio and television, which are top-down and linear (Royston & Creeber, 2009, p. 142). The Habermasian public sphere calls for dialogic, deliberative, communicative and democratic ideals (Royston & Creeber, 2009, p.

142), and the interactivity has the potential to reach these ideals. Online communication can approximate real-life verbal exchanges that serve as the basis for the Habermasian public sphere (Royston & Creeber 2009, p. 142). Paschal Preston (2001) argues that there are problems with the democratic ideals of the digital public sphere, as Internet applications can "only provide an opportunity for an open and interactive public sphere" (cited in Royston & Creeber, 2009, p. 142). There is also a possibility for people taking advantage of digital media. Also, the dominant commercial and political interests in society can take advantage of the new media. Last, he points out the problem of technological literacy (Royston & Creeber, 2009, p. 142), meaning that all of the population should be equally able to use the Internet and digital applications in order for its members to reach these communicative ideals.

Next, we would like to look into the special characteristics of the Ukrainian public sphere. According to Freedom House's analysis of Freedom on the Net (2018), the Ukrainian Internet is partly free after a sharp decline in Internet freedom in 2017, when Ukrainian authorities blocked pro-Russian or pro-Separatist webpages. Also, in June 2017 there were cyber-attacks on the country. Internet penetration in the country is 52.2% of the population of 44,831,159 (Freedom House, 2018). To understand the scale, it could be compared to, for example Finland, where the Internet penetration of the population is 75% (Tilastokeskus, 2018).

The relationship with Russia is an important aspect in the digital public sphere of Ukraine. The conflict in Crimea and the information war with Russia pose challenges to Internet freedom in Ukraine: it is vulnerable to cyber-attacks and misinformation campaigns, which leads to a debate on Internet regulation. There has been a demand for more regulation (Albrecht, 2019). In addition to Russian influence, oligarchs dominate the political scene. Oligarchs are powerful, well-connected businessmen who have divided up the country's economic assets among themselves and use government funds for their own profit (Oldhauser, 2014, p. 92). However, Ukrainian civil society has a strong presence online and activists have used social media to reach several goals, for example, to organise volunteer support for military functions, stay current with developments in the east of Ukraine, and promote human rights; and additionally, they have revealed biased and manipulated information online. There has been a proposal to tighten Internet freedom by the means of laws to control the digital public sphere, but until now such laws have not passed because of the pressure from local civil society (Freedom House 2018). Despite the pressure, debate on digital rights is minimal and the protection of personal data is weak (Albrecht, 2019).

If we look into the ideals of a public sphere that according to Habermas are dialogic, deliberative, communicative and democratic, and look into the digital public sphere that would be an interactive, lateral and discursive public sphere, how would we characterize the Ukrainian public sphere on this scale of ideal – non-ideal? First, the interactivity of the Internet is a factor that no amount of regulation can take away, unless the whole Internet system is shut down, since interactivity is essential to the Internet and the era of the digital public sphere. This makes it possible to have deliberation in the public sphere not only in a "top-down" way but also through interaction between the "top" and the "bottom", as in the case of presidential election between politicians and the citizens. This interaction is making the digital public sphere more ideal by letting voices at the grassroots be heard.

On the other hand, Internet regulation, the power of the organizations, people or the government, can make it harder for those voices to be heard. As noted earlier by Preston, the ideal is only an ideal and an opportunity, but there are several aspects that can make it harder for the public sphere to reach its ideals. One aspect that he mentions is the possibility of people taking advantage of the digital public sphere, as commercial and political interests in society can have an effect on it (Royston & Creeber, 2009, p. 142). Here we can see that these kinds of challenges for the public sphere could be, for example, the two characteristics of Ukrainian society we mentioned earlier: Russian influence and oligarchs. Russian influence (political interest) disturbs information flow itself by the way of cyber-attacks and misinformation campaigns for its own benefit. Oligarchs could be seen as the economic power holders, who interfere in the flow of information by gaining benefit for themselves through politics, and therefore naturally through media, also in the digital public sphere.

Finally, we would like to point out the effect of Internet penetration in Ukraine. We have already stepped into the era of the digital public sphere, since the digital media of the current era, such as television and the Internet, play an important role in public discussion. Even though the Internet and social media are important, television still has its place in the digital public sphere. Television has a huge impact on the formation of public opinion in Ukraine, since for 85% of the population it is the main source of information (Nieczypor, 2018).

To some extent, Ukraine is living up to the standards of a deliberative, communicative and democratic public sphere, since the Internet is partly free. A strong online presence of civil society ensures the realization of the ideal of the public sphere to some extent, since people can communicate and have deliberation through the interactive relations. Perhaps the digital public sphere in Ukraine could be called a kind of semi-democratic digital public sphere.

## 3. Fragmented public sphere and the 2019 Ukrainian presidential elections

The fragmentation of the public sphere has been a topic of discussion for theorists for a long time (e.g. Arendt, 1958) and Habermas (1989) developed the theoretical investigation. Currently this fragmentation is a topic that divides theorists, as some claim that the fragmentation has a negative effect on society, e.g. leading to a loss of the common good or common interest among people (McKee 2005, p. 141) and weakening public engagement and debate (Hodkinson 2017, 192). Some scholars argue that fragmentation threatens deliberative democracy, because people should be exposed to opposing opinions (Downey & Fenton 2003, p. 185). On the other hand, however, some argue that the fragmentation of the commonly shared public sphere has worked as a power against homogeneity and allowed e.g. different minority voices to be heard by the common public, as well as new issues to be brought forward for public discussion (McKee 2005, p. 146).

Habermas himself sees the fragmentation of the public as a destructive phenomenon. He questions whether the public sphere worked better earlier, before, for example, women and the working class were allowed to vote or participate in public society. This is not because he opposes human rights but because the women and the working class brought their own stories, discourse and matters to the public sphere, thus fragmenting the homogenous public society (McKee, 2005, p. 145). From the Habermasian standpoint, it seems that the fragmentation of the public sphere was and is a historical challenge and is not solely based on the development of the Internet (which is a commonly analysed factor in the studies of fragmentation today). However, as the research on the fragmentation of the public sphere has evolved since Habermas and he has been criticized for basing his views on the patriotic, bourgeois public sphere of the 18th and early 19th centuries (Hodkinson, 2017, p. 196), his views might not be the most applicable tool for analyzing modern phenomena.

The fragmentation of the public sphere has been a topic of research since at least Donald Trump’s victory in the 2016 United States presidential election, since his victory took a large part of the population by surprise. However, not everyone finds this fragmentation alarming. According to McKee, some theorists claim that the public sphere has always been fragmented and different issues have always divided the citizenry into competing interest groups (McKee, 2005, p. 142). The media landscape itself also adds to the fragmentation and to the dividedness of society with the media outlets diversifying into smaller units to respond to different needs (Hodkinson 2017, p. 192). The Internet is also adding to this as it is facilitating the divide into smaller and smaller interest groups as users wish to only discuss certain topics or issues. (Papacharissi, 2002, p. 17).

Another issue that is closely related to the fragmentation of the public sphere is the echo chamber phenomenon. Echo chambers are not necessarily created deliberately as a political marketing strategy but might be a product of natural human coping mechanisms (Zakharchenko et al., 2019). Some theorists see the whole Internet as an echo chamber. It has been discovered that people tend to look for opinions that are similar to their own, and the phenomenon is explained by theories such as cognitive dissonance and selective exposure theories. This way the Internet does not necessarily expose users to new information, different opinions or opposing political views, but rather reaffirms the political orientation an individual already holds (Arvidsson et al., 2014). It is possible that a fragmentation of the public sphere has always existed, but the Internet effectively works as an echo chamber, reinforcing already existing political views.

How can these theories of fragmentation be applied to the Ukrainian public sphere? The digital sphere became an important factor in Ukraine during the protests around 2013. Since then the topic also captured the attention of researchers, who found that the circumstances increased the political and social importance that the digital sphere had in Ukraine (Zakharchenko et al., 2019) and Zelensky also used that in his campaign, which was mostly conducted online. It has also been discovered that Zelensky used different campaign messages to different groups of people (Ben, 2019). The echo chamber phenomenon and fragmentation throughout the Internet might add to the fact that different interest groups do not necessarily have any contact with each other, and Zelensky was able to use this to his advantage in his campaign.

At a first glance Zelensky could be seen as a unifying power in society and the Ukrainian public sphere especially compared to many other presidents if we compare the Ukrainian results to those in countries with a strong conservative-liberal divide in elections, such as the Trump election or the Brexit vote. Judging from that framework, gaining an immense majority the way Zelensky did could even be seen as a unifier of Ukrainian society and public sphere. He gained 73% of the votes (Ben, 2019) which could signal that much of the population, fragmented public sphere aside, desired change and Zelensky managed to brand himself as one to deliver change.

Looking more closely at Zelensky voters supports this claim, as e.g. in the case of NATO, Zelensky voters were rather divided (37% supported joining NATO and 37% were against it), yet most of these people voted for him. This could be seen as proof of the earlier argument about Zelensky as unifying the Ukrainian public. However, Ben has also discovered that different campaign messages were targeted at different people. This strategy shows Zelensky and his team possibly understanding the concept of fragmentation in the public sphere and using it cleverly to their advantage. This is where further fragmentation of the public sphere in society could pose a threat or raise issues, because different interest groups not communicating with each other enables this to happen. If a candidate is able to target different groups with different (or, quite frankly, opposing) campaign messages (without being called on it), it is questionable whether a common public sphere even exists.

However, some research points to the direction that the state of the Ukrainian public sphere contributed to the possibility of a populist leader gaining power in the Ukrainian election. The demand for a new political leader and the overall dissatisfaction among the citizenry were not answered in the common public sphere that was largely controlled by the oligarchs in Ukraine. Therefore, the fragmentation of the public sphere into smaller entities was essential for the strategy of Zelensky's campaign and the grounds for his popular support (Haran & Burkovsky, 2019).

It needs to be addressed that the reasons behind Zelensky's victory include most likely many other factors besides the fragmentation of the public sphere. Research already conducted on the election underlines the importance of populism and the typical (unrealistically) easy solutions to complex issues, as well as the empty discourse typical of populists. In Zelensky's case the voters were ready to accept all of his promises at once (Dodonova, 2019). In addition, Zelensky's campaign consisted mostly of going against and defaming the incumbent Petro Poroshenko. Zelensky tried to profile himself as a unifier of society and as working against the fragmentation—even his slogan was "Servant of the people". Due to the huge majority of Zelensky voters, the election could be seen as a unifying factor in the Ukrainian public sphere.

## 4. Zelensky's entirely virtual presidential campaign in the digital public sphere

The successful 2019 presidential campaign of Ukraine's Volodymyr Zelensky is a particularly representative case study of questions relating to a 21st century digital public sphere and its manifestation in real world politics. The whole campaign and landslide electoral victory were both an astonishing case of life imitating art and an example of certain features of a digital public sphere in concrete electoral politics. The completely virtual campaign is also an illustrative example of both the participatory potentials and shortcomings of the digital media environment as a contemporary public sphere. The current section analyses Zelensky's campaign from this perspective, discussing how it showed emancipatory potential akin to the participatory liberal model of a public sphere, but came short in challenging features upholding to deliberative democracy. The section ends with a discussion of how due to the democratic limitations of new media technologies also inherent here, the campaign may have had more in common with what Papacharissi (2002, 2010) calls a public space instead of a public sphere.

In his campaign Volodymyr Zelensky cast himself as a total political outsider, but in fact he had literally performed politics in the public eye for quite some time, as the lead actor in the immensely popular television series Servant of the People (2015-2019), in which he plays a common and honest man who inadvertently gets elected president of Ukraine and subsequently works to purify the state's corrupt political system. Capitalizing on the ready image of his likable and popular fictional character, Zelensky's campaign was truly a campaign of the modern era, because it was practically completely virtual.  During the four months between the announcement of his candidacy and election day in April 2019, he "did no face-to-face campaigning, made no speeches, held no rallies, eschewed travel across the country, gave no press conferences, avoided in-depth interviews with independent journalists and, until the last day of campaigning, did not debate" (Karatnycky, 2019). What Zelensky did do was attempt to communicate more directly on his own terms with the electorate through videos on his YouTube channel, posted on his social media accounts, appeared on television shows.  Apart from this he also engaged people by calling on them to virtually give him advice in designing his political platform (Karatnycky, 2019).

Zelensky's upset victory against established political institutions in Ukraine has been described as a people's rebellion, "a rebellious popular vote against corrupt politics-as-usual by … [a heterogeneous spectrum of] voters [lacking] any political unity or ideological coherence", unified only by a general distaste of the status quo (Cherepanyn, 2019). In this sense, it can be argued that Zelensky's campaign fulfilled a participatory liberal model of the public sphere, which can be described as a "perspective on democracy that particularly stresses the benefits of active engagement in politics both for the citizen as an individual and for the system as a whole" (Ferree et al., 2002, p. 299). Above all else, participatory liberalism emphasizes mass mobilization of citizens and ideas, striving for "the widest possible empowerment, and popular inclusion is necessary to achieve this" (p. 299). It can be argued that Zelensky's campaign catered to precisely these demands, as his "virtual-first strategy allowed him to run his campaign on general themes and vague promises and to avoid issuing detailed positions on policy issues … [instead h]is political messaging focused on discontent with the way things are" (Karatnycky, 2019). In a sense, Zelensky's campaign, fictional character and even political party named after the television show, all acted as an empty signifier, a blank slate for a very wide range of different identities and demands among citizens of a fractured country, whose main priority was to achieve a change to the status quo, but without a unified vision of what the contents of that change should be.

Such a promise of change was offered by Zelensky with his broad, imprecise platform that nevertheless reached people and engaged them directly through the digital media environment. The campaign that took place in and of the digital public sphere worked primarily as a platform that facilitated a reaction to hegemonic politics. In that sense, the outcomes were emancipatory, in the sense of participatory liberal theory of the public sphere. On the other hand, Zelensky's platform was so vague that while a reactionary emancipation of the voice of the people was achieved, it seems to have offered little in the way of a platform for deliberative democracy in the digital public sphere. Indeed, when compared to the discursive model of the public sphere, which "shares the value of popular inclusion with participatory liberalism, but unlike that tradition, views this as a means to a more deliberative public sphere rather than as an end in itself" (Ferree, et al., 2002, p. 306), such deliberativeness is difficult to detect as a feature of this "people's rebellion".

Incidentally, the shortcomings of the participatory liberal tradition vis-à-vis the discursive tradition identifiable in this case study are reminiscent of what concerned Habermas himself when he spoke of the public sphere's structural transformation and "criticized that mass media, and commercial mass media in particular, do not further deliberation" (Schäfer, 2015, p. 2). Optimists have attempted to lay these concerns to rest by pointing to the 21st century transformation of the public sphere into one taking place in cyberspace, framing the Internet "as the very arena in which global civil society will come together to forge public opinion and facilitate collective action" (Delacruz, 2009, p. 10). However, problems with this argument are evident in the case of Zelenskiy's campaign, which demonstrated to great extents the digital features of the contemporary public sphere, but also its limits in that the process seemed to fall short of being conducive to deliberative democracy. Instead it seemingly empowered the electorate but did not set deliberativeness and equitable consensus-building at the center of this empowerment, which played out mostly for the sake of protest.

These shortcomings of the digital public sphere are illustrated by Papacharissi (2010), who argues that while new digital technologies have the potential to revitalize the public sphere, they do not inherently lead to greater degrees of deliberative democracy. Indeed, she argues that access to information and communication channels does not automatically equate greater civic engagement, connections made online are often characterized by mistrust instead of trust, and commercialization of the Internet often sets economic motives above democratic ones (pp. 121-123). The above-mentioned features of digital technologies are ones that create a public space, but are not as conducive to the creation of a genuine public sphere (p. 124). Thus, Ukrainian empowerment is still limited to something like the campaign of Zelensky, which is a very interesting contemporary illustration of the possibilities inherent in the practical workings of the digital public space aspiring to be a digital public sphere. However, as it played out, such a campaign is one that accommodates only an ambiguous set of demands and “serves as a mere screen for the popular political imagination” (Cherepanyn, 2019), instead of going a step further and facilitating digital deliberative democracy.

## 5. The influence of the digital public sphere in the Ukrainian presidential election and its effects

Worldwide, the Internet has changed the way people communicate and interact. Citizens, candidates, and parties have been using social networks and different types of technology to run their campaigns. Over the last 20 years, the role of information and communications technologies (ICTs) in election campaigns has evolved from purely support tasks like mailing, graphical design, and database, to new direct channels of communication between political parties and candidates (Chen & Smith, 2010, p. 4).

Nowadays, the Internet had opened a window to the creation of content and free communication between users of the network. The emergence of social media platforms such as Facebook, Twitter, YouTube and blogs, among others, has allowed the world to communicate efficiently in different levels and contexts. Users can freely publish content and spread their ideas and opinions without physical interaction. As Çela observed, "It was never as easy as it is now for the people to come together and be organized to express their criticism or to contradict a certain matter that concerns a certain community” (2015, p. 196).

On the other hand, while users have more presence on the Internet, they are more exposed to several types of content such as products, information, and general advertisements (Stephen, 2016, p. 7). These types of content, added to consumer behavior and marketing research, are being used by specialized companies to create targeted digital strategies to influence the decisions and choices of users. Even the political public sphere has been affected by digital platforms. For instance, during the election of political candidates, digital campaigns have taken on an essential role in the final results. There have been cases where citizens' information was used to deliver political propaganda based on their preferences to influence them. However, not all of the information is reliable. It has been showed that some fake news was delivered purposely.

Grossi (2011, p. 2) in his article, "The public sphere and communication flows in the era of the Net," said, "It is undeniable that the transformations of the public sphere in late modernity societies are increasingly interlinked with the growing pervasiveness of the Net—both in the flows of top-down political communication and bottom-up discursive practices which also make public alternative issues, multiple belongings, and new rights". In that regard, the communication process has transformed with numerous constituencies, and vertical and horizontal communication channels.

Politicians can directly interact with their followers and receive feedback immediately without intermediaries. The gap between these two actors of society has become a one-on-one dialogue. Therefore, the digital public sphere is considered as an important societal change, because the digital world can be a place for the renewal for public political discourses (Thimm, 2015).

Over time candidates have used traditional advertisement to run their political campaigns. However, during the last decade the massive surge of social networks has allowed politicians to take use as tools to run their campaigns. For Chen & Smith (2010, p. 4) the employing of digital channels has "increased professionalisation and reintegration of online channels into the core marketing strategies". A clear example of that is the Americans presidential elections of 2008 and 2012, when former President Obama's campaign used digital platforms like Twitter, Facebook, and Snapchat. And in the last American presidential campaign, Donald Trump's team used citizens' information data to predict fundraising and digital political advertising (Benkler, Faris, & Roberts, 2018, p. 271).

Numbers are solid when we look at the public sphere's influence in the recent presidential campaign in Ukraine. Social networks were the main platforms Volodymyr Zelensky used as a strategy to communicate his message to the citizens across the country. In the social networks VK, Facebook, and Instagram, the Servant of the People party got more reactions and mentions in posts than Petro Poroshenko's European Solidarity party, Zelensky's opponent. (Matviyishyn, Iliuk, & Panchenko, 2019).

Some people have "compared the Zelensky presidential campaign to U.S. President Donald Trump, who won the 2016 elections as an anti-elite candidate with a business background and an appeal to protest voters tired of established politicians" (Talant, 2019). However, evidence suggests that the Ukrainian campaign was more similar to that of former President Barack Obama's, because Zelensky communicated more directly to the electorate through social platforms to communicate with his supporters.

As mentioned before, social networks allowed citizens and politicians to communicate and express their opinions with each other. In the case of Ukraine, Zelensky understood that the best way to articulate his ideas was using new channels. He had a strong presence on Facebook and Instagram. The latter had 18,759 posts using the hashtag #ServantOfThePeople during the campaign in comparison to the opposite party. European Solidarity, Poroshenko's party had 1024 posts using the name of the party as a hashtag (Matviyishyn et al., 2019). It is important to refer to the numbers because they help us see the impact of social networks.

According to Internews Ukraine (2019), VK, Facebook and Instagram are used for different purposes, with different views and opinion leaders. The Servant of the People party had a major presence on Instagram. VK content is dominated by Russian narratives, and Facebook is more diverse in its political narrative. Therefore, social networking tools create links and channels for a better, direct, and cohesive exchange of information and communication between the government and the citizens, instead of the traditional channels, such as television and printed propaganda.

Another clear platform used by the Servant of the People party and Zelensky himself was YouTube. Applying exceptional and excellent video production and filming several videos, he was able to present himself to the people of Ukrainian as a fresh new face in politics. He even gained popularity because the traditional media (TV news) showed his videos. There is no doubt that Zelensky used digital campaign as the main mechanism to gain support.

We can also see the presidential elections in a perspective of marketing and consumer behavior. If we look at Zelensky's presidential campaign as a marketing strategy, we can apply the observation that social networks are often platforms to sell products. In this case, politicians sell ideas and political proposals to citizens. Thereby, consumer behavior could be influenced by the information that is being delivered to them. The "Servant of the people" party certainly utilized the best tools to influence voters and to reach new audiences.

The issue of the digital public sphere that features ubiquity, user-generated content, multimediality, and portability has become a key issue for the process of mediatisation (Thimm, 2015). The changes in social media communication can affect and create socio-cultural transformation in society. In the Ukrainian context Zelensky was able to win the elections using new ways of communication, generating sociocultural changes in politics.

Certainly, the digital public sphere influenced the way elections run in Ukraine. The strategies Zelensky and his party used in their campaign are a precedent for future elections in Ukraine and around the world. Traditional media, such as TV, radio, and news still have a significant influence, but the digital public sphere is transforming communication, consumer behavior, politics, and society.

Ukraine's digital public sphere will continue to evolve. The government and politics will face new challenges in the upcoming years. Citizens will persist in changing the flow of the public sphere and new audiences will emerge to interact in the communication process. Politics will have to adapt to sociocultural changes in new global contexts.

## Conclusion

In this essay we analysed the digital public sphere and the Ukranian presidential elections of 2019. Whilst the concept of the public sphere and phenomena related to it are not by themselves enough to completely explain the election result, they do work as a sufficient tool for analysis in this essay.

The first section examined how the concept of the public sphere has been extended in Habermas' theoretical works. After a contrast with the argument by Arendt, on which Habermas bases the notion of the public sphere, we argued that from *Structural Transformation of the Public Sphere* (1962/1989) to *Between Facts and Norms* (1992/1996), he modified the concept in an expansive way by departing from the historical particularity which constitutes the basis of the original bourgeois public sphere articulated in his early work and by putting more emphasis on versatility in the modern world. Thus, we argued that the normative elements of his concept are still applicable to our analysis of the recent Ukraine case.

The digital public sphere in Ukraine has its own special features that might explain some functions in the public sphere and how the presidential campaign was implemented and received. The Ukranian public sphere is partly controlled by Internet regulation, the power of organizations and people, but there is also a strong online presence of civil society. Besides the online presence, television still holds an important position in the formation of public opinion. These structural features cannot be neglected when researching the implications of the digital public sphere in the Ukranian context.

The fragmentation of the public sphere alone as a theory is not sufficient to explain Zelensky's victory. However, some of his campaign methods used this fragmentation to his advantage. The fragmentation of the public sphere might be to blame for enabling a political candidate to send contradicting campaign messages to different groups of people. On the other hand, the considerable majority of the votes that were cast for him could signal the existence of a rather widely shared public opinion, which could be interpreted as a sign of a common public sphere.

Zelensky's campaign was able to use this fragmentation to its benefit largely thanks to its wholly virtual character, through which it was possible to keep the message imprecise on actual policies but still tap into a common resentment felt by Ukrainians, for varying reasons, toward the current political status quo. This method made it possible for large swathes of the electorate representing very different political identities to back Zelensky. The result of this revolt was the realization of an emancipatory potential of the digital public sphere, but due to the thinness of the policies discussed, a more deliberative democratic potential was not realized.

The digital public sphere certainly influenced the Zelensky presidential elections in a way that led him and his party to the presidency. Even though this is not the first example of using ICTs such as digital platforms and social networks in political campaigns, it is the first time that the whole campaign was conducted in the digital public sphere. Digital platforms and new technologies will become a guiding force in the future of politics shaping and influencing global political processes.

## References

Adorno, T. W. & Horkheimer, M. (1947/2002). Trans., Jephcott, E. *Dialectic of Enlightenment: Philosophical Fragments*. Stanford, CA: Stanford University Press.

Albrecht, E. (01.07.2019, accessed 4.11.2019) Could digital freedom in Ukraine become a casualty of Russia's info Wars? *Deutsche Welle*. https://www.dw.com/en/could-digital-freedom-in-ukraine- become-a-casualty-of-russias-info-wars/a-49264448

Arendt, H. (1958). *The Human Condition*. Chicago: The University of Chicago Press.

Ben, B. (2019). Why Zelenskyi Won. Euromaidan Press. Retrieved from http://euromaidanpress.com/2019/04/23/why-zelensky-won-ukraine-presidential-elections/

Benkler, Y., Faris, R., & Roberts, H. (2018). Mammon's Algorithm. In *Network Propaganda* (pp. 269–288). Oxford University Press.

Burkovskiy, P. & Haran, O. (2019) Ambivalence of public opinion as the base for populism in the 2019 presidential campaign. Головна сторінка eKMAIR. Accessed 4.11.2019. http://ekmair.ukma.edu.ua/bitstream/handle/123456789/16374/Ambivalence_of_public_opinion_as_the_base_for_populism.pdf?sequence=

Calhoun, C. J. (1992). *Habermas and the Public Sphere.* Cambridge, Mass.: MIT Press.

Çela, E. (2015). Social Media as a New Form of Public Sphere. *European Journal of Social Sciences Education and Research*, *4*(1), 195. https://doi.org/10.26417/ejser.v4i1. p195-200

Chen, P. J., & Smith, P. J. (2010). Adoption and Use of Digital Media in Election Campaigns: Australia, Canada and New Zealand Compared. *Public Communication Review*, *1*(1), 3. https://doi.org/10.5130/pcr.v1i1.1249

Colleoni, E., Rozza, A. & Arvidsson, A. (2014). Echo Chamber or Public Sphere? Predicting Political Orientation and Measuring Political Homophily in Twitter Using Big Data. *Journal of Communication 64*(2), 317-332.

Delacruz, E. M. (2009). From Bricks and Mortar to the Public Sphere in Cyberspace: Creating a Culture of Caring on the Digital Global Commons. *International Journal of Education & the Arts 10*(5), 1-21.

Dodonova, V. (2019). Tandem of Populism and Post-Truth as the Background of Development of the Modern Democracy in Ukraine. ResearchGate. Retrieved from https://www.researchgate.net/publication/334406333_Tandem_of_Populism_and_Post-truth_as_the_background_of_development_of_the_modern_democracy_in_Ukraine

Downey, J. & Fenton, N. (2003). "New media, counter publicity and the public sphere". New Media and Society. SAGE Publications. Retrieved from https://journals.sagepub.com/doi/10.1177/1461444803005002003

Ferree, M. M., et al. (2002). Four models of the public sphere in modern democracies. *Theory & Society, 31*(3), 289-324.

Freedom House. (2018, Accessed 4.11.2019) "Freedom on the Net 2018" https://freedomhouse.org/report/freedom-net/2018/ukraine.

Grossi, G. (2011). *The public sphere and communication flows in the era of the Net**. (2005), 1–34. Retrieved from http://www.ciberdemocracia.net/victorsampedro/wp-content/uploads/2012/12/Grossi.pdf

Habermas, J. (1962/1989, 1991). Trans., Burger, T. *The Structural Transformation of the Public Sphere: An Inquiry into a Category of Bourgeois Society*. Cambridge: Polity Press.

———— (1984, 1987). *The Theory of Communicative Action: Reason and the Rationalization of Society*, Vol. 1 and Vol. 2. Boston, Massachusetts: Beacon Press.

———— (1992/1996) Trans., Rehg, W. *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy.* Cambridge: Polity Press.

Hodkinson, P. (2017) *Media, Culture and Society*. London: SAGE Publications.

Internews, U. (2019). Social networks and the election in Ukraine. What Facebook, Instagram, VK users are saying about the parliamentary race. Retrieved from Internews Ukraine website: https://internews.ua/en/opportunity/fb-instagram-vk

Karatnycky, A. (2019, April 24). The World Just Witnessed the First Entirely Virtual Presidential Campaign. *Politico*.

Martin, R & Creeber, G. (2009, accessed 4.11.2019) *Digital Cultures*. Maidenhead: McGrew-Hill Education.

Matviyishyn, I., Iliuk, O., & Panchenko, M. (2019). Memocracy: What role social networks played in Ukraine's parliamentary vote. Retrieved from https://ukraineworld.org/articles/infowars/memocracy-what-role-social-networks-played-ukraines-parliamentary-vote

McKee, A. (2005). *The Public Sphere.* Cambridge: Cambridge University Press.

Nieczypor, K. (28.02.2018) Serving Politics. Television's role in Ukraine's presidential election. Centre for East Asian Studies (OSW). https://www.osw.waw.pl/en/publikacje/osw-commentary/2019-02-28/serving-politics-televisions-role-ukraines-presidential

Oldhauser, E. (2014) (edit.) *Democratization in Ukraine, Georgia, and Belarus: success, stagnation*. Happauge: Nova Science.

Papacharissi, Z. (2002) The Virtual Sphere. New Media and Society. SAGE Publications. Retrieved from https://journals.sagepub.com/doi/10.1177/14614440222226244

———— (2010) *A Private Sphere: Democracy in a Digital Age*. Cambridge: Polity Press.

Schäfer, M. S. (2015) Digital Public Sphere. In G. Mazzoleni, et al. (Ed.), *The International Encyclopedia of Political Communication* (pp. 322-328). London: Wiley Blackwell.

Stephen, A. T. (2016). The role of digital and social media marketing in consumer behavior. *Current Opinion in Psychology*. https://doi.org/10.1016/j.copsyc.2015.10.016

Talant, B. (2019). Inside Zelenskiy's campaign: How social media, TV fame can win him presidency. *Kyiv Post*. Retrieved from https://www.kyivpost.com/ukraine-politics/inside-zelenskiys-campaign-how-social-media-tv-fame-can-win-him-presidency.html?cn-reloaded=1

Tilastokeskus. (4.11.2018, accessed 4.11.2019) The Internet is used ever more commonly with a mobile phone – even for shopping. https://www.stat.fi/til/sutivi/2018/sutivi_2018_2018-12- 04_tie_001_en.html

Thimm, C. (2015). Citizen Participation and Political Communication in a Digital World. In A. Frame & G. Brachotte (Eds.), *Citizen Participation and Political Communication in a Digital World*. https://doi.org/10.4324/9781315677569

Wilson, M.I. & Corey, Kenneth E. (2012) "The role of ICT in Arab spring movements". Networks and communication studies. *Digital Territories: Case Studies 26*:3/4 p. 343-356. Accessed 4.11.2019. https://journals.openedition.org/netcom/1064?lang=en

Zakharchenko, A., Maksimtsova, Y., Iurchenko, V., Shevchenko, V. & Fedushko, S. (2019) "Under the Conditions of Non-Agenda Ownership: Social Media Users in the 2019 Ukrainian Presidential Elections Campaign." *Ceur-ws*, *2393*. Retrieved from https://arxiv.org/ftp/arxiv/papers/1909/1909.01681.pdf

# Part III

# Environment

# 3.1 Encountering Sustainable Communication and Greenwashing: Environmental Values in Organizational Communication

Julia Ylä-Outinen, Mikaela Rydberg, Annina Mattila, Lisa Kärnä, Anni Helkovaara
Faculty of Social Sciences, University of Helsinki

# Abstract

In this study we examine how different organizations communicate their commitments to sustainability and corporate social responsibility on their websites, and the different ways stakeholders could interpret this communication. We do this by examining several case studies and reflecting on those cases with the help of a theoretical framework. Our main findings are that there is a growing concern amongst stakeholders regarding environmental values and that unsubstantiated sustainability claims issued in corporate publicity can be interpreted as greenwashing. We identify a conflict between goals of growth versus environmental sustainability in some of the cases. We also discover that organizations appear to be more transparent with their intentions by communicating their environmental values based on firm-serving motives rather than public-serving ones.

*Keywords*: Sustainability, organizational communication, environmental communication, greenwashing, corporate social responsibility, aviation, food industry, fashion industry, sustainability

In this research paper we aim to examine how different organizations communicate sustainability and corporate social responsibility on their websites and the different ways stakeholders could interpret this communication. We also aim to investigate what makes claims of sustainability believable. We chose to examine five organizations from several fields of business, that play a significant role in the everyday life of western consumers, while also being some of the greatest global contributors of carbon emissions and greenhouse gases.

In this text we refer to terms related to both communicative methods and sustainability questions. We discuss the terms greenness and corporate social responsibility. The concept of greenness and green is usually used when referring to products or production that takes environmental aspects into account. "Green" is associated with closeness to nature and environmental characteristics (Polonsky, 1994). Corporate social responsibility (CSR) is referred to in many parts of our research and can be interpreted in various ways depending on the case. Kotler and Lee describe CSR as "a business' commitment to contribute to sustainable economic development, working with employees, their families, the local community, and society at large to improve their quality of life." (2005, p. 3)

There are many reasons for why recognizing environmental aspects and reporting CSR is now more important to organizations than ever. Noticing the limits of the globe's carrying capacity and the more rapid spread of information through traditional media, as well as new social media channels, are among these reasons. We begin this research paper by examining two organizations from the textile industry, Hennes & Mauritz and Patagonia. Then we examine two organizations from the food industry, Valio and Oatly. In the last section we examine air traffic through looking at Finnair. Finally, we present our findings and conclusions.

## The textile industry

Today's textile industry is based on an enormously polluting model of fast fashion: the clothes are bought, used a couple of times and then thrown away. The British Ellen McArthur Foundation (2017) has calculated that 73 per cent of the clothes produced globally end up burned or in a landfill, instead of the materials being reused.

According to the foundation's research paper, today's textile industry is dominated by and optimized for cotton and polyester. Currently, polyester makes up 55 percent and cotton 27 percent of total textile fiber production. The global fashion and design company Hennes & Mauritz features prominently in the Ellen McArthur Foundation's research.

### H&M

The purpose of this case study is to find out how global fashion and design company Hennes & Mauritz's (H&M's) marketing portrays responsibility and compare it to existing research on organizations' environmental and sustainability discourse. How does the company portray environmental values in their online communication to consumers? I will focus on examining the communication and marketing messages on the company's website and in its 2018 Sustainability Report. Additionally, I ask what the organization actually does to fulfil its environmental promises.

I chose to look at H&M because they have been accused of greenwashing by the media and environmentalists several times. H&M is and has been a part of several sustainability projects and has made promises to improve their sustainability practices. The purpose of this case study is to find out how H&M portrays sustainability on its website, and whether one can find instances of greenwashing in H&M's sustainability marketing.

The H&M Group is the second-largest global clothing retailer after Inditex. H&M Group owns the brands H&M, Cos, Monki, Weekday, Cheap Monday and & Other Stories. Net sales of the group were SEK 210 billion in 2018. According to their website, H&M has approximately 4900 stores in 73 markets and 50 markets with online shopping.

### Corporate social responsibility as a concept

Corporate social responsibility, or CSR, is "a commitment to improve community well-being through discretionary business practices and contributions of corporate resources" (Kotler & Lee 2005, p. 3). This means that an organization's actions can be classified as responsible if the organization has committed to operating on behalf of the environment.

Stakeholders are actors that the organization interacts with and is dependent on (Juholin 2017, 142). In the eyes of the stakeholders, CSR is more and more strongly associated with the organization's legitimacy (Pollach, Johansen, Nielsen, ja Thomsen, 2012, p. 205). Especially for big organizations, corporate social responsibility is now a critical part of their strategy. However, besides just communicating commitment, the organization needs to actually be responsible, as well.

### Criticism

H&M has failed to deliver on its corporate social responsibility related promises before. In 2013, the company announced that all "H&M's strategic suppliers should have pay structures in place to pay a fair living wage by 2018. By then, this will reach around 850,000 textile workers" (Washington Post 26.11.2013). The deadline of 2018 passed and no information was shared on whether or not the company actually succeeded in delivering the promise. In May 2018, the international campaign #TurnAroundHM was started. The campaign criticised H&M's unfulfilled promise and demanded that the company publishes a clear plan of action, making sure that the sewers get fair living wages.

H&M has also received criticism on the way it collects used clothes. This can be interpreted as greenwashing, because the consumer may think it is okay to buy new clothes now that they have returned the old ones.

### H&M on sustainability on their website and the 2018 Sustainability report

H&M states that in 2018 the company's customers handed in 20,649 tonnes of old textiles as part of the company's program for reuse or recycling. The company also says that they used 96 percent renewable electricity in their own operations in 2018. However, there is no mention of how many clothes the organization produces in a year.

On the company's "About Us" page the first headline states that H&M is committed to Ethical and Sustainable AI. They also have a "Sustainability" headline that has subtitles called "People", "Planet" and "Sustainability Reporting". Because the purpose of this text is to examine the sustainability reporting through environmental glasses, the section under "Planet" is what I examined.

Under the subtitle there are six articles: materials, recycling, climate, circularity, chemicals and water. In total, H&M offers a lot of information on what ways it plans to be or already is sustainable. Key notions include that according to H&M's Annual report 2017, H&M Group has a commitment to use 100 per cent renewable energy in its own operations. The share was 96 per cent in year 2017. According to the report, H&M Group has also committed to becoming climate positive throughout its entire value chain by 2040 at the latest.

On H&M's website, we also find the company's 2018 sustainability report. It is a 208-page-long report that has some figures and strategy explained. However, a great deal of the report is "vision" and "innovation" rather than actual numbers.

### Greenwashing as a concept vs. H&M on sustainability

Pollach, Johansen, Nielsen and Thomsen have argued (2012, p. 207) that corporate social responsibility can both be seen as contributing to reputation and stakeholder relationships, and the organization's actual operation. If the CSR is only a contributor to reputation and not an actual part of the company's operations, or if the company exaggerates its green intentions, this action may be interpreted as greenwashing. Greenwashing means "misleading consumers about their environmental performance or the environmental benefits of a product or service" (Delmas & Burbano, 2011).

The importance of reputation has grown. Juholin (2017, p. 49) presents concepts of image and reputation. The concept of image was generated in the beginning of the 1900's and it means creating desirable mental images through communicative ways. Reputation is a newer concept and it is based on actions rather than mental images.

One fact that supports the claim that H&M greenwashes its operations is that the company switches its clothes collections every season. On their website, the company claims to aim for a circular business model wherein "resources stay in use for as long as possible before being regenerated into new products and materials, resulting in a reduction in waste and negative impacts". Yet at the same time the company replaces an old collection with a new one every season. The goal is to sell as many clothes as possible.

Another contradiction related to the recycling service is how H&M states in the sustainability report that "all Monki customers are rewarded with a "10% off your next purchase" voucher when they bring a bag of unwanted textiles for the garment recycling service." The same kind of offer (a discount or a free voucher) is used in the H&M Group's other stores. The purpose is to get the customer to come to the store again to buy more products but H&M describes it as "a reward for sustainable behavior" (H&M Sustainability report 2018, 29).

### Conclusion

Corporate social responsibility is quite a new trend among organizations and reporting it has become more popular in recent years (Mäkelä and Kujala 2017). The research on transparency is also more public than before. Seventy-five percent of consumers have said that their purchasing decisions are influenced by a company's reputation with respect to the environment, and eight in ten have said they would pay more for products that are environmentally friendly (Klein 1990). This means that some organizations may claim to be responsible because they actually want to be, whereas others might just do it for the profit and for gaining a good reputation.

## Patagonia

The California-based company Patagonia makes clothing and equipment for the outdoors. With a strong focus on sustainability and corporate responsibility, they are often cited as a leading company when it comes to sustainability issues. Having based much of their communication on these matters, Patagonia is a good example for this study. In the following text I aim to map out what kinds of sustainability communication strategies Patagonia uses and how the consumer might react or respond to these. My main argument is that a company that has sustainability at its core can successfully and believably communicate and place arguments about its products' "greenness" to consumers. There is, of course, a paradox in Patagonia's communication. While Patagonia might represent "greenness" and eco-friendliness to many, people are also consuming and "buying into" the process of the fast-growing clothing industry while making these purchase decisions.

### Theoretical Framework

In order to understand and interpret communication by Patagonia, it is necessary to apply communication theory. Godemann and Michelsen (2011) outline what could be seen as communication about sustainability. They point out that there are often certain methods that are used to influence and manage the process of communication (2011, p. 9). These methods include social marketing, empowerment, instruments of participation and planning or education (2011, p. 9). Through examining Patagonia's website and social media channels, the easiest and most accessible sources of information, all of these methods can be identified.

Godemann and Michelsen (2011) point out the practice of orienting communication to a specific audience or people with certain lifestyles about social issues such as sustainability. The social marketing approach is, according to Godemann and Michelsen, a vital communicative method today, because word-of-mouth communication and the web have increased their meaning (2011, p. 9). This method has proven to be effective when campaigning for voluntary actions and supporting behavioral changes (2011, p. 9). Patagonia shows signs of these kind of strategies in their campaign "The New Localism". "The New Localism" is a campaign consisting of different places that need protection and attention. Patagonia phrases this inclusive campaign through wording like "We are all locals" and "we all have a chance to make a difference" (a. patagonia.com "New Localism", accessed 16.10.19). Patagonia also emphasizes the importance of in- and out-groups in their stakeholder communication. Michael Polonsky points out in his article how companies are beginning to realize their position as members of a wider community when it comes to sustainability and responsibility. Polonsky states how this shift often results in a corporate culture that integrates these values (2005, 5). When examining Patagonia's corporate culture and the communication related to this, it is very clear that sustainability is at the core of corporate culture and that this culture is heavily built around a framework of sustainability and "teamwork" in the fight against global warming (b. Patagonia.com, Employee activism, accessed 16.10).

Another strategy for sustainability communication is empowerment. Godemann and Michelsen describe this strategy as an enabling process where participation is key and self-assessment plays a vital role (2011, 9). According to Godemann and Michelsen this happens in practice through workshops and other participative methods (2011, p. 9). When it comes to self-reflection, Patagonia seems to be very thorough. A big part of their corporate communication consists of self-reflective analysis (c. Patagonia.com "The Responsible Company", accessed 16.10.19). They openly speak of their early setbacks and mention what still needs to be done in order to become a fully responsible company. Patagonia admits to the "twin conundrums" as they call it (d. patagonia.com "the Shell Game in the Dark", accessed 27.10) when campaigning for new products in a market where growth is the main focus.

The last method of sustainability communication that Godemann and Michelsen mention is examining educational processes (2011, p. 10). Through learning processes that create autonomous action instead of just trained behavior, real behavioral changes can be made. Godemann and Michelsen mention how processes in education for sustainable development can be sparked by both formal and informal educational actors (2011, p. 10).

Patagonia does not emphasize its role as an educator in sustainability issues, although they provide a vast amount of information for consumers on their website. The company speaks openly of possible challenges and success stories, painting a picture of an active learning process. “We need everyone in the fight, so we share proprietary information and best practices with other businesses, including direct competitors. Our business is a tiny fraction of the global apparel industry, and we know we can’t solve the climate crisis alone. We also know we don’t have all the answers.” (e. patagonia.com “Activism”, accessed 27.10.19). The tone of voice in Patagonia’s communication can be seen as humble but determined as well as encouraging (Fowler and Hope, 2007, p. 32). Many of Patagonia’s posts on social media channels encourage followers to take action. These encouragements include getting in touch with congress, taking part in climate strikes and urges for consumers to learn more. This gives the reader a possibility to study multiple themes and aspects concerning materials, factories and corporate social responsibility.

Fowler and Hope point out in their in-depth study how Patagonia, in a pioneering project of using new materials, withdrew from the project at the last minute (2007, p. 33). The company devoted two pages of their catalogue and large billboards to explain this turn of events to their stakeholders (2007, p. 33). It is interesting to view this openness in decision making as an open learning process in which all stakeholders can feel included. Being sincere about both success as well as challenges and concerns may provide extra value in communication for the consumer. Peter Dauvergne and Jane Lister point out in their book, “Eco-Business: A Big-Brand Takeover of Sustainability,” how consumers have become increasingly suspicious and less trusting. According to Dauvergne and Lister, consumers want to gain more influence and feel more involved in what they buy and what products they are offered (2013, p. 131).

Dauvergne and Lister mention how companies are embracing corporate sustainability and how they can maximize profits though this “strategic CSR” (2013, p. 4). This “eco-Business” as they call it, is an example of how business advantages can be made through environmental efforts (2013, p. 4). Dauvergne and Lister mention how communicating sustainability issues can be interpreted as added value by the potential customer. Sustainability communication is, according to Dauvergne and Lister, a way to differentiate products and a way to drive increased sales (2013, p. 122).

## Conclusion

All this leaves us with the dilemma of “complicated greenness” presented by Sharon Hepburn (2015). Patagonia’s marketing strategies and business ideas revolve around sustainability and encouraging people to appreciate and preserve the Earth. By doing this the company also encourages consumption over the actual needs of their customer (Hepburn, 2015). Patagonia has tried to distance themselves from the world of fashion through “minimalist style” and through enhancing “simplicity and utility”

(f. Patagonia.com, "Mission statement" accessed 28.10). This holistic approach to sustainability issues combined with the encouraging and informative communication flow creates a trustworthy and wholesome communicative environment.

## The food industry

The food industry has a significant impact on our environment through, among other things, greenhouse gas emissions, the use of land and water resources, pollution and the impact of chemical products. Studies show that the food industry is responsible for 20-30% of the environmental impact of private consumption. Within the industry, meat and meat products have the greatest environmental impact. The second greatest impact is created by dairy products, followed by a variety of others, such as plant-based food products, soft drinks and alcoholic drinks. (EIPRO 2006.)

Many organizations ranging from The United Nations (2018) to Greenpeace (greenpeace.org, "#Food" accessed 27.10.2019) have called for a rethinking of dairy and meat production and a decrease in consumption in order to tackle climate change.

### Valio

More milk is consumed in Finland than in any other nation in the world: 361 kilograms per person a year. Milk production is the main source of income in rural areas of the country. Eighty-five percent of Finnish beef is sourced as a byproduct of milk production. (The Natural Resources Institute Finland, 2019).

In this research paper I examine how Finnish milk and dairy producer Valio Oy communicates sustainability and corporate social responsibility on its website. I will begin by giving a brief introduction of Valio. Then, I will present some of the key ideas and concepts related to this section. Next, I'll take a look at Valio's website and dissect the ways in which accountability and corporate responsibility are present. Finally, I will reflect on my findings.

Valio Oy is the largest producer of milk and dairy products in Finland. It has a dominant position in the industry, and it exports its products to over 60 countries. On its website the company reported a turnover of 1,734 million euros in 2018. Valio is owned by Finnish milk producers through cooperatives.

I chose to explore Valio as a case study in part because in recent years the company has faced some less than positive publicity. In 2018 Valio had to pay small dairy producers 8 million euros in fines, after it was caught pushing its pricing under the market value (Yle 18.6.2019). In 2019 the Finnish corporate responsibility organization Finnwatch revealed several human rights violations in Valio's supply chain in Thailand (Yle 26.3.2019). Valio has also been in the headlines recently, when Helsingin Sanomat criticized Valio for claiming its milk comes from free range cows — even though one third of those cows still live chained up in cowsheds (Helsingin Sanomat 6.9.2019).

## Sustainability communication, green marketing and greenwashing

Sustainability communication is found throughout the public sphere. Its goal is to tackle sustainability issues such as climate change, the shrinking of biodiversity and consumption. (Godemann and Michelsen 2011, 10). It is beneficial for a corporation to communicate sustainability and green values to consumers and stakeholders because of the growing concerns over climate change. This is where green marketing comes in. Peattie and Charter define green marketing as a holistic approach that is both sustainable and profitable. It highlights global concerns and does not treat the environment as a means to an end (Peattie and Charter 2003, p. 727). According to Tinne (2013, p. 1), green marketing is the marketing of products that a consumer can assume are environmentally safe. By definition, green marketing has to be backed up by the organization's sustainable acts. If no such acts are present, the organization may be guilty of greenwashing. According to Jenner, "greenwashing consists of any advertising, marketing or public relations actions by corporations to project an image of being an environmentally-minded organizations, even when their business practices are destructive" (2009, p. 9). Greenwashing is not a new issue, and the term has been used since the 1980's to describe false claims of sustainability (Dahl 2010, p. 247). Green marketing and signaling sustainable values can be perceived as greenwashing by consumers if they feel there is a discrepancy between the organization's words and its actions.

## Sustainability and corporate social responsibility on Valio's website

Valio highlights corporate social responsibility and sustainability on its website. The first picture on the website is a large banner for Valio's promise to reduce its carbon footprint to zero by 2035.  Images of nature, happy looking cows ranging on sun-dappled fields and smiling farmers are frequent.

There are several links to articles about sustainability on the front page. All of the articles state the same core mission of reducing the company's carbon footprint to zero. In these articles Valio states that it is going to give up fossil fuels step by step, replace its plastic packaging with recycled plastic and convert fully to using soy-free cow fodder. They also name "accountability" as one of their core values.

Valio's website has a whole section for accountability and corporate social responsibility. This section covers several topics ranging from animal welfare and sustainable milk production to transparency and wellness innovations. There is also an "Accountability Report" for the year 2018. This report states the goals mentioned earlier, and includes numbers and graphs containing information about Valio's emissions and energy consumption at the end. However, the report contains very little concrete evidence of actions Valio has already taken in order to become more sustainable.

### Conclusion

Milk and dairy production, and meat as its byproduct, are amongst the biggest contributors to greenhouse emissions in Finland, causing 12% of the country's greenhouse emissions (Statistics Finland 2018, p. 12). Valio likes to highlight its greenness and accountability. On its website the company covers all the topics that are usually brought up when discussing sustainability of the dairy industry: animal welfare, use of soy in fodder and carbon and methane emissions.

Valio acknowledges the issues the industry faces on its website. Based on its sustainability goals, it seems that the company's claims of accountability are somewhat backed up by its actions: giving up soy fodder is one example of those actions. On the other hand, it is important to take note of the fact that much of the company's sustainability is still in the policy phase. Thus, for example Valio's goal of being carbon neutral by 2035 is not yet backed up by any proof. It is beneficial for corporations to publish environmental policies, because they can have a positive influence on public opinion, market share and stakeholder relations (Ramus & Montiel 2005, p. 378). However, if these policies aren't realized, this type of green marketing can also be perceived as greenwashing, as there is no evidence backing the claims of sustainability.

If we follow the definition of greenwashing by Jenner (2009), presented earlier in this research paper, it is necessary to wonder whether a company like Valio can ever communicate its greenness without being guilty of greenwashing. When we consider the effects of dairy and meat production and consumption on global climate change, it seems like the only sustainable option is to scale back on both. As a business, Valio's goal is growth. As long as this is the case, it does seem like the company's claims of sustainability are just a Band-Aid on a bullet hole.

De Vries, Terwel, Ellemers and Dancker (2013, p. 143) state that, "regardless of the company's intentions, in the end it is all about whether or not people perceive corporate greenwashing". I believe this to be very important in the case of Valio. Even though most of the company's sustainability goals haven't yet been realized, the Finnish public seems to believe in the company's accountability: Valio has won the Sustainable Brand Index B2C Finland six times in a row, most recently in 2019 (Sustainable Brand Index Official Report, 2019). It appears that the public does not see a conflict between Valio's brand, actions and sustainability policy. It will be interesting to see if the recent negative media coverage has any impact on the company's reputation and the public's perception of its sustainability and accountability.

## Oatly

In the following section I focus on a Swedish company called Oatly. I analyze the marketing communication on their website and focus on how they talk about sustainability and green values in their communication.

Oatly produces oat-based products. The company's patented enzyme technology copies nature's own process. It turns oats into nutritional liquid food and offers an alternative for dairy based products. Oatly values sustainability as their goal is to offer plant-based drinks that are in tune with the needs of both humans and the planet (Oatly 2019a.). By offering an alternative for dairy products, Oatly also offers an alternative for the dairy industry, which can impact the environment in various ways, the scale of that impact depending on the practices of the dairy farmers and feed growers (WWF 2019).

## Key concepts

With the planet facing increasingly serious environmental problems, green consumption and sustainability have been gaining more attention by consumers and companies alike. Companies develop their green marketing strategies to show their customers social responsibility and their good corporate image (Zhang, Li, Cao, Huang 2018). However, sustainability is not only related to the environment. Rather it is about finding some sort of balance so that Earth can support the human population and economic growth without ultimately threatening the health of humans, animals and plants. The elements that form the basis for sustainability are environment, economy and equality. It is argued that sustainability can be achieved by simultaneously protecting the environment, preserving economic growth and development, and promoting equality (Portney 2015).

Jacob Vos (2009, p. 681) introduces some of the most common forms of corporate greenwashing. He names environmental policy statements as a usual form of greenwashing. With a broad, high-minded statement proclaiming a corporation's commitment to preserving the environment, the statements often make an impression of an environmentally friendly company. The picture the statements paint is often highly idealistic. The policy statements rarely include any specifics regarding the implementation of the policy. Without specific commitments, the statement is not bound to any benchmarks which might be used to measure its progress toward its goal.

Generally speaking, people suspect less strategic behavior when a company communicates an economic motive for investing in environmental measures. Studies also show that companies that express firm-serving (economic) motives are seen as relatively trustworthy (Zhang et al. 2018). Research has also shown that organizations that communicate public-serving motives are considered less honest. That in turn provokes less trust towards the company than towards organizations that communicate organization-serving motives (Terwel, Harinck, Ellemers, Daamen 2009).

## Sustainability and environmental marketing on Oatly's website

Oatly claims to be "not just another company" that only sells products. Rather they frame the company as one with an ideology based on sustainability, health and transparency (Oatly 2019a). Oatly’s marketing communication on their website is open-hearted, approachable and simplified. They aim is to show their social responsibility by saying things as they are and by being transparent in their marketing communication. They showcase their products with lists of the origin of all the ingredients as well as links to the supplier websites. The product pages also include a description of their product and what is environmentally good and not so good about them. By showing the customers the company’s commitment to social responsibility, they also assess their work and what could be done more sustainably.

Oatly promises to be a good company (Oatly 2019), but they do acknowledge that it is impossible to be completely good. “We are not a perfect company, not even close, but our intentions are true. We would like to be judged by the good we do and not just the pretty words we say” (Oatly 2019). This exemplifies the transparency in Oatly’s marketing communication. They acknowledge that green values are an integral part of the company and not only a marketing tool they use. But the company’s positions of being “green”, and the environmental measures they take serves their economic goals as they are a company producing plant-based goods.

Environmental sustainability is not only about reducing carbon emissions and reliance on fossil fuels. It is also about other natural resources: most notably water and land. (Portney 2015). Oatly highlights the carbon emissions of their products but doesn’t forget these other aspects of sustainability. Ways to make land usage more sustainable are especially discussed in their sustainability report.

In their marketing communication, Oatly mainly uses firm-serving rather than public-serving motivations. Oatly’s whole business in centered around plant-based products. In the recent years the global need for climate-friendly products has been growing rapidly (Oatly 2019b). Oatly however was founded back in the 1990s and has been offering these climate friendly products well before they became a global trend. By positioning themselves as an environmentally friendly company and endorsing green values, Oatly expresses firm-serving motivations as their business is based on offering an alternative for dairy products.

### Oatly’s sustainability report

A closer look at Oatly’s sustainability report shows that there are concrete goals that the company has set for themselves. In addition to introducing these goals, the sustainability report also shows how the company plans to achieve them, what risks are involved, what their current status is in achieving each goal and where the reader can get further information about the subject (Oatly 2019b). By setting the goals as well as showing the reader what is being done to achieve those goals, Oatly emphasizes their commitment to their environmental policy and underlines the fact that the company is not only “talking the talk, but also walking the walk”.

Perhaps the most challenging aspect of sustainability is the “equality” element. The importance of equality seems to lean on the fundamental assumption that an unequal world is an unsustainable world. Moreover, conceptual work on sustainability hasn't made entirely clear how equality relates to the economic and environmental elements of sustainability (Portney 2015). In their sustainability report, Oatly mainly treats the subject of equality as a gender related question rather than taking a wider view of the subject.

### Conclusion

Environmental values and sustainability are a core part of Oatly's operations. The marketing communication on their website reflects their sustainable values by being transparent. Oatly heavily communicates firm-based motivations. Environment issues and sustainable values are an integral part of their company's values.

Considering that environmental motives are probably not the only reason for companies to invest in environmental measures, it seems advisable from a strategic perspective to be reticent in claiming purely selfless motives in public communications in order to avoid being perceived as greenwashing (Vries, Terwel, Ellemers & Daamen 2015). If a company wants to market environmental values, they should also show what being environmentally friendly brings to the company, to appear more honest. Oatly's transparency makes their marketing communication appear more honest. Environmental values and sustainability are not only framed as a central part of the company's marketing communication but also appear integrated as inseparable parts of Oatly's business model.

## Aviation

Aviation contributes to around 2% of the world's global carbon emissions. An economy-class return flight from London to New York emits an estimated 0.67 tonnes of CO2 per passenger, according to the International Civil Aviation Organization (ICAO, 2019). Flights burn fuel and produce greenhouse gases, which in turn contribute to global warming by releasing carbon dioxide into the atmosphere. As flying is becoming more and more frequent, the carbon footprint of the airline industry has been under tight scrutiny.

### Finnair

Global passenger numbers have been predicted to double to 8.2 billion by 2037 (ICAO, 2019), which would mean a rapid increase in air emissions. As environmental problems have entered public discussions, airline companies have been forced to rethink their corporate social and environmental responsibility. In this paper I focus on one airline company, Finnish government-controlled Finnair.

### Key concept

As Lynes and Andrachuk state in their research "Motivations for corporate social and environmental responsibility: A case study of Scandinavian Airlines", the driving force of what makes companies commit to social and environmental issues can be unpacked into internal, sector-specific and external influences (Lynes & Andrachuk, 2008, p. 378). Catalysts help shape influences by acting as a medium for encouraging or discouraging corporate social and environmental responsibility. These catalysts could be for example the financial position of the firm, internal leadership within the firm as well as the culture the firm operates in. From an environmental point of view, strongly influential catalysts seem to be the culture the firm operates in (Lynes & Andrachuk, 2008, p. 380).

## Sustainability and the aviation industry

The air travel industry has been moving towards a highly competitive phase, wherein market pressure has been lowering prices and promoting the introduction of more efficient and competitive products aimed to serve consumers from all economic backgrounds (Pilling, 2004, p. 46).

Airline flights within Europe are covered by the EU's emissions trading system (ETS), which provides the worst emitters with an incentive to reduce their carbon pollution. (European union emissions trading system, 2019). Each year airline companies have to surrender a number of permits equivalent to the amount of carbon dioxide they emitted in the preceding year. Permits are acquired through an annual allocation system and some are issued for members for free. If polluters don't have enough allowances to acquit their previous year's emissions, they can buy additional permits at auctions or from other companies. The EU has put a maximum cap on the CO2 emissions that can be emitted by restricting the number of permits available. Emitters are thus provided with an incentive to reduce their emissions, because this is cheaper than buying scarce permits (European union emissions trading system, 2019).

The negative image associated with the environmental impacts of air travel also pushes airline companies to be more socially and environmentally responsible corporations. Airline companies also represent their home countries, and as flag carriers of the country, national airline companies have a certain responsibility to uphold a positive image of their country of origin (Clancy, 2001).

## Findings on Finnair's sustainability report

I examined Finnair's sustainability reports, which state their motivations in corporate social and environmental responsibility. Finnair proclaims on its website that it is committed to building a more sustainable aviation industry with its stakeholders. It has a dedicated page dedicated to corporate responsibility, where consumers can use an emission calculator and get to know the company's environmental values. It is also possible to offset the CO2 emissions of one's flights or reduce them by buying biofuel for the aircraft.

In 2018 Finnair paid 11 million euros for the EU's emissions trading system permits, when the company's whole revenue was almost 3 billion euros (Yle, 8.4.2019). At the same time Finnair was given a not-so-flattering ranking as being one of the biggest emissions growing airline companies in Europe. Finnair's emissions from flight operations grew by 11.9 percent in 2018 compared to the previous year, making its total emissions around 3.2 million tonnes of CO2. When comparing only the carbon dioxide emissions, Finnair is Finland's second biggest company to impact global warming.

On its website Finnair does not mention being one of the worst CO2 emitters in Finland, but it does address its growing emissions. In 2018 Finnair's jet fuel consumption increased by 109.6 million kilograms, approximately 11.9 percent compared to the previous year. Finnair explains this by traffic growth and operational challenges.

This traffic growth can be explained by the company's new flight routes to Asia, that cover two thirds of the company's carbon dioxide emissions. But this does not explain all of the growing emissions, since Finnair's emissions grew also in the European markets. In the company's sustainability report there is a table that breaks down the emissions into smaller sections and compares the 2018 emission numbers to those of the past three years. This table shows that in 2018 Finnair didn't use any renewable jet fuel at all, even though it now encourages customers on their website to compensate for their flights by buying biofuel.

Finnair's fuel efficiency, as in its capacity to use fuel according to payload weight, was 251 revenue tonne kilometers in 2018. This is slightly higher than the 2017 figure (247.6 RTK), but less than in 2016 and 2015 (both 271.2 RTK).

## Nordic values are important in Finnair's brand image

In the sustainability report Finnair states, it is going to renew its fleet. According to Finnair, a modern fleet is going to reduce flight emissions. The introduction of these new technologies that involve cleaner productions and lower production costs underpin one of Finnair's main motivations with respect to environmental commitment: financial cost-benefits of environmental management. Finnair is committed to achieving carbon neutral growth by 2020 and cut emissions of its flight operations by half by 2050 from the 2005 level.

Throughout the report it is possible to see the Nordic culture which Finnair as a company works in. References to Nordic values can be found in the text, as well as pictures of clean Finnish forests and winter landscapes. As Finnair is Finland's national airline company, the outlook of the report and the branding of the firm go hand in hand with Finland's brand image of being a clean, Nordic country that is proud of its forests.

This image that the website and the sustainability report create highlights the importance of the culture the firm is working in. It determines its social and environmental motivations; thus, the ideology of Nordic responsibility and cleanliness has a profound influence on Finnair's environmental commitment.

### Conclusions of the sustainability report

Reading Finnair's sustainability report it seems that the airline company is conscious of its responsibility as a high emission producing company. It appears it has already started to improve its actions towards more sustainable choices, but it is not clear whether an airline company can ever be sustainable and environmentally friendly. Flying is almost portrayed as an ecological action as customers are encouraged to replace their carbon footprint by buying biofuel for the aircraft. This leads to the question whether if biofuel is really that much of an improvement, why isn't Finnair willing to switch all of the jet fuel it is using to biofuel and compensate the price on the flight prices? Now the choice of using biofuel is left to the customer, which creates the impression that the environment is only important to the company if it is important to the stakeholder.

It is also questionable whether these kinds of choices that are left to customers are actually doing anything for the company's sustainability. This merely seems to shift the moral responsibility to the customers, without the company taking any responsibility of its own.

## Conclusion

In this research paper we examined how different organizations communicate sustainability and corporate social responsibility on their websites. We also examined whether their claims of sustainability could be interpreted as believable by stakeholders.

Our main findings can be stated as following:

- There is growing concern amongst stakeholders regarding environmental values and sustainability issues. As studies have shown, 75 percent of consumers say that their purchasing decisions are influenced by a company's reputation in respect to the environment.
- Green marketing and sustainability communication can be interpreted as greenwashing by stakeholders if there is a conflict between an organization's words and actions. Communicating sustainability can therefore damage an organization's reputation and even hinder its growth when it is not backed up by actions.
- A conflict with corporate growth can be identified in many of the cases. As companies grow and environmental actions are put on a pedestal, there is pressure to expand and develop, sometimes at the expense of greenness.
- As the need for environmentally friendly products is growing rapidly, organizations appear to be more transparent with their intentions by communicating their environmental values based on firm-serving motives rather than public-serving motives.

As a final thought we would like to point out the conflict of overall consumption. As environmental questions are constantly on the surface of communication, in both a corporate context and in the media, we hope to see more focus put on sustainability questions and consumerism. As stated earlier, holistic sustainability communication combined with concrete actions and evidence is considered believable by stakeholders. Yet, actions and responsibility must be a shared effort.

# References

Clancy, M. (2001). Exporting paradise: tourism and development in Mexico (p. 127). Amsterdam: Pergamon.

Dahl, R. (2010). Green washing. Environmental Health Perspectives, Jun; 118 (6): 247-252

Dauvergne, P. (2013). Eco-business: A Big-brand Takeover of Sustainability. Cambridge, Mass.: MIT Press.

De Vries, G., Terwel, B. W., Ellemers, N. and Daamen, D. (2015). Sustainability or Profitability? How Communicated Motives for Environmental Policy Affect Public Perceptions of Corporate Greenwashing. Corp. Soc. Responsib. Environ. Mgmt. 22: 142–22, 142–154.

Delmas, M. A., & Burbano, V. C. (2011). The Drivers of Greenwashing. California Management Review, 54(1), (pp. 64–87). https://doi.org/10.1525/cmr.2011.54.1.64

Ellen McArthur Foundation Report (2017). Retrieved from https://www.ellenmacarthurfoundation.org/assets/downloads/publications/A-New-Textiles-Economy_Full-Report_Updated_1-12-17.pdf

Environmental Impact of Products (EIPRO). (2006). Analysis of the life cycle environmental impacts related to the final consumption of the EU-25 Main report. Retrieved from https://ec.europa.eu/environment/ipp/pdf/eipro_report.pdf.

European Union Emissions Trading System. (2019). Retrieved from https://ec.europa.eu/clima/policies/ets_en#tab-0-1

Finnair. (2019). Retrieved from https://company.finnair.com/fi/corporate-responsibility#Cleaner

Finnair. (2019). Sustainability report. Retrieved from https://company.finnair.com/resource/blob/1353302/a5d7eae9c5038a568a614004cb56f1fb/finnair-sustainability-report-2018-data.pdf

Fowler, S. J. (2007) “Incorporating Sustainable Business Practices into Company Strategy." Business Strategy and the Environment 16, no. 1: 26-38. doi:10.1002/bse.462.

Godemann, J. Michelsen G (2011). Sustainability Communication: Interdisciplinary Perspectives and Theoretical Foundation. Dordrecht; New York: Springer.

Greenpeace. (2019). Retrieved from https://www.greenpeace.org/international/tag/food/

Hepburn, Sharon J. (2013). “In Patagonia (Clothing): A Complicated Greenness." Fashion Theory 17, no. 5: 623-645. doi:10.2752/175174113X13718320331035

H&M. (2019). Retrieved from https://www.hm.com

H&M’s Annual report 2017 [https://about.hm.com/content/dam/hmgroup/groupsite/documents/en/Digital%

20Annual%20Report/2017/Annual%20Report%202017%20Sustainable%20development.pdf].

Jenner, E. (2005). Greenwashing: visual communication and political influence in environmental policy. Louisiana State University. LSU Doctoral Dissertations. 3887.

Juholin, E. (2017). Communicare! Viestinnän tekijän käsikirja. Helsinki: Inforviestintä.

Klein, J. T. (1990). Interdisciplinarity: History, Theory, and Practice. Detroit, MI: Wayne State University Press.

Kotler, P. & Lee, N. (2005). Corporate Social Responsibility: Doing the Most Good for Your Company and Your Cause.

Lynes, J. K., & Andrachuk, M. (2008). Motivations for corporate social and environmental responsibility: A case study of Scandinavian Airlines. Journal of International management, 14(4), 377-390

Milk production. (2019). The Natural Resources Institute Finland. Retrieved from

https://www.luke.fi/en/natural-resources/agriculture/milk-production/

Mäkelä, H. & Kujala, J. (2017). Integroitu raportointi yritysvastuun mittaamisen ja arvioinnin näkökulmasta. Teoksessa Juholin, E. & Luoma-aho, V. (toim.) Mitattava viestintä: ProComma Academic 2017. Helsinki: ProCom – Viestinnän ammattilaiset ry. (pp. 110─121)

Mäki, M. (2019, June 18). Valio tuomittiin maksamaan miljoonakorvaukset paikallismeijereille – Yhtiö myi perusmaitoa alihintaan. Yle. Retrieved from https://yle.fi/uutiset/3-10836876

Nalbantoglu, M. (2019, October 27). Valio kertoi, että sen tölkkimaidot ovat jatkossa vapaan lehmän maitoa – kolmasosa Valion lehmistä elää silti yhä kytkettynä parteen. Helsingin sanomat. Retrieved from https://www.hs.fi/talous/art-2000006230358.html

Oatly - About Oatly (2019a). Retrieved from https://www.oatly.com/fi/about-oatly.

Oatly - Sustainability report 2018 (2019b). Retrieved from https://www.oatly.com/uploads/attachments/cjzusfwz60efmatqr5w4b6lgd-oatly-sustainability-report-web-2018-eng.pdf.

Patagonia - New Localism. (2019a).  retrieved from https://eu.patagonia.com/fi/en/new-localism.html

Patagonia - Emplyee Activism (2019b). retrieved from https://eu.patagonia.com/fi/en/employee-activism.html

Patagonia - The Responsible Company". (2019c). retrieved from https://eu.patagonia.com/fi/en/responsible-company.html

Patagonia - A Shell Game in the Dark. (2019d). retrieved from https://eu.patagonia.com/fi/en/a-shell-game-in-the-dark.html

Patagonia - Activism. (2019e). retrieved from https://eu.patagonia.com/fi/en/enviro%2C-environmentalism%2C-environmental-activism%2C-patagonia-environmentalism%2C-activism%2C-action-works%2C-patagonia-action-works%2C-our-voice/our-voice.html

Patagonia - Patagonia's Mission Statement". (2019f). retrieved from https://eu.patagonia.com/fi/en/sustainability.html

Peattie, K. & Charter, M. (2003). Green Marketing. The Marketing Book (5): 726-755.

Pilling, M. (2004). Brand Extensions, Airline Business, 20(3), pp. 46-48

Pollach, I., Johansen, T. S. Nielsen, A. E. & Thomsen, C. (2012). The integration of CSR into corporate communication in large European companies. Journal of Communication Management, 16 (2), (pp. 204─216)

Polonsky, M. J. (1994). "An Introduction to Green Marketing." https://helka.finna.fi/PrimoRecord/pci.escholarshipqt49n325b7.

Portney, K. E. (2015). Sustainability. Cambridge, Massachusetts: The MIT Press.

Ramus, C. A., & Montiel, I. (2005). When Are Corporate Environmental Policies a Form of Greenwashing? Business & Society. 44(4): 377–414.

Sustainable Brand Index. (2019). Retrieved from https://www.sb-index.com/finland

Suomen kasvihuonekaasupäästöt 1990-2017. (2018). Statistics Finland. Retrieved from http://www.stat.fi/tup/julkaisut/tiedostot/julkaisuluettelo/yymp_kahup_1990-2017_2018_19735_net.pdf

STT. (2019, March 26). Finnwatchin selvitys: Valion tavarantoimittajat polkevat siirtotyöläisten oikeuksia Thaimaassa. Retrieved from https://yle.fi/uutiset/3-10706812

Terwel, B. W., Harinck, F., Ellemers N. & Daamen, D. (2009). How organizational motives and communications affect public trust in organizations: The case of carbon dioxide capture and storage. Journal of Environmental Psychology 29 (2): 290-299. doi 10.1111/j.1539-6924.2009.01256.x

Tinne, W. S. (2013). Green Washing: An Alarming Issue. ASA University Review. 7 (1): 81-88.

United Nations Environment Programme. (2018). What's in your burger? More than you think. Retrieved from https://www.unenvironment.org/news-and-stories/story/whats-your-burger-more-you-think

Vastuullisuusraportti 2018. (2018). Valio. Retrieved from https://ejulkaisu.grano.fi/valio/Vastuullisuusraportti2018#p=1

Vos, J. (2009). Actions Speak Louder than Words: Greenwashing in Corporate America. Notre Dame Journal of Law, Ethics & Public Policy 23 (2): 673-698. doi 10.33531/farplss.2019.3.04

Vries, G., Terwel, B. W., Ellemers, N., & Daamen, D. (2015). Sustainability or Profitability? How Communicated Motives for Environmental Policy Affect Public Perceptions of Corporate Greenwashing. Corp. Soc. Responsib. Environ. Mgmt. 22: 142– 154. doi 10.1002/csr.1327

Washington Post. (2013). H&M says it will pay factory workers a "fair living wage." It doesn't say what that means. Retrieved from https://www.washingtonpost.com/news/wonk/wp/2013/11/26/hm-says-it-will-pay-factory-workers-a-fair-living-wage-it-doesnt-say-what-that-means/

WWF - Sustainable Agriculture: Overview. (2019). Retrieved from https://www.worldwildlife.org/industries/dairy.

Yle. (2019). Lentoliikenteen kasvun nurja puoli: Finnairin päästöt kasvoivat eniten Suomessa, yhtiö yksi eniten päästöjään lisännyt lentoyhtiö EU:ssa. Retrieved from https://yle.fi/uutiset/3-10725918

Zhang, L., Li, D., Cao, C. & Huang, S. (2018). The influence of greenwashing perception on green purchasing intentions: The mediating role of green word-of-mouth and moderating role of green concern. Journal of Cleaner Production 187: 740-750. doi: 10.1016/j.jclepro.2018.03.201

# 3.2 Governing Climate Change

Anni M. Taskinen, Roosa M. Savo, Leo H. Pahta, Roosa M. Kontiokari,
Dongyang Huo
Faculty of Social Sciences, University of Helsinki

## Abstract

Climate change is a multi-faceted, complex phenomenon. It cannot be narrowed down only to the changes occurring in our natural world, as it is also a social phenomenon that affects societies, cultures, and politics both at the macro- and microscopic levels. The complexity of both physical climate change as well as its socially constructed aspects, combined with the urgent need for climate action, pose a great challenge for sustainable global climate politics and climate governance. Climate governance means the “mechanisms and measures aimed at steering social systems toward preventing, mitigating, or adapting to the risks posed by climate change” (Stripple & Sverker 2003, 388).

In this paper, we aim to explore the construction of climate change as a social phenomenon and discuss various approaches to governing climate change, such as deliberative democracy, applications of game theory, discourse and algorithmic governance.

*Keywords:* Climate change, deliberative democracy, game theory, algorithmic governance, imagined communities

## 1. On the construction of climate change as a social object — L.H. Pahta

While climate change is a highly visible topic in the current political and societal debate, knowledge of the phenomenon itself is not new. The scientific basis for the notion of human industry influencing Earth's climate has been understood since the 19th century: The idea that an increased concentration of carbon dioxide has an effect on rising global temperatures was first proved by T. C. Chamberlin in 1896 (The White House, 1965).

While the science behind climate change was firmly established by scientists such as Chamberlin, it took almost another century before serious governmental regulation to curb the fast-growing emissions was finally enacted. This had to do with the growing sense of urgency of the topic, caused both through advances in measuring increases of CO2 in the atmosphere, as well as through increased public understanding of the level of devastation the warming planet would bring for all humans.

Rachel Carson's Pulitzer-prize winning book *Silent Spring* was published in the year 1962 and is widely regarded as the springboard from which modern climate change discussion begun (The New York Times, 2012). Still, climate change emerged on the global environmental governance agenda only in the late 1980s as a threat second only to nuclear war (Bentley, 2017). Unfortunately, we are still working towards nations globally taking the required steps towards curbing global warming (D'Aprile, 2018).

The discourse in which climate change has been handled in public discussions has changed during the years (Ylä-Anttila et al., 2018). What the combination of all the discussions on the climate during the years have established, is gradually forming a social object — one worth discussing on a global scale, and one on which we are all expected to have an opinion on. In other words, the physical phenomenon has been politicized.

To help us find solutions to mitigate climate change, it helps to understand what this social object entails and how it is formed. This understanding gives us capabilities to find new ways of acting in a more sustainable manner, to form governance to establish common rules of conduct, and to finally effect change to curb global warming.

Social objects are placeholders for patterns of activities (Searle, 1995). The term social object refers to objects, such as institutions, that gain their meaning through processes of reification. We believe money to be money through the accepted common and continuous use of the objects imbued with this meaning. The same thing can be said of climate change, as a physical phenomenon and as a social object we assign a function to.

Understanding climate change requires expert knowledge of several complicated systems of interdependencies. The Earth's climate is formed through a combination of natural phenomenon, which interact with each other in extremely complicated ways. To help form an epistemic viewpoint of the climate, we can use tools to measure the changes which occur and form historical series to put these changes in context. Through careful analysis, we can then ascertain that the yearly weather cycles around the world are in fact changing. Finally, through understanding that human industries emit carbon dioxide combined with a physical understanding of how particles act in a closed system, we can deduce that humans play a part in climate change.

This function that we assign to climate change is non-agentive, that is, it helps form a theoretical account of the phenomenon in question but isn't part of the actual physical phenomenon. As we live in the physical world though, it is unavoidable that a changing climate has an overarching effect on our lives. What is left to debate is how we should react to this fact. Thus, climate change enters our social reality as a problem to be tackled through various forms of human intervention.

According to Downs (1973, p. 39), public debate follows a certain issue-attention cycle, which is rooted "both in the nature of certain domestic problems and in the way major communications media interact with the public". In this model, after the first big splash a news item makes, the reaction becomes more muted as more and more new topics become newsworthy.

While the news cycle of our times is increasingly unrelenting with breaking news stories appearing continuously on news pages, news also has the power of building narratives and forming agendas for the public discussion (Scheufele 2006, p. 9). These discussions continue and mutate in social media platforms after the news themselves have disappeared from feeds. This reification, continuous acting out of climate change as a global problem on a scale of impending doom on social media is what makes the phenomenon as a social object more real. It also directs more and more individuals towards action in some capacity in a community formed through climate change.

During the past few years we have seen the rise of several social movements working towards pushing nations to slow down the rise of global temperatures and halt them at the level of 1.5 Celsius compared to 1990 levels (IPCC, 2019). These include Fridays for Future, and Climate Strike and Extinction Rebellion. These organizations have managed to gather hundreds of thousands of people on the streets to demand climate justice.

In Benedict Anderson's germinal book *Imagined Communities*, he speaks of communities being imagined into being by people who perceive themselves as part of a certain group (Anderson, 1983, p. 6). While Anderson's book dealt mainly with how nationalism is constructed, we can see this development also in the case of climate change discussion, where the discussion breeds political activity and loyalties to movements.

The idea of human-based climate change allows for communities to form by finding various strategies towards finding solutions to stop it, bringing together various people, who without this idea might not have worked together. It allows for protesting, demanding change, finding and developing new, less energy-intensive ways of living, being more mindful of one's own actions towards generating waste, and shaping one's everyday life towards a more sustainable lifestyle. It is only through pronouncing that humans have an existential problem in the form of climate change, that we can begin to constitute it as a social artifact, upon which we can direct our actions.

Despite the clear scientific results of humans' effect on the climate, public discussions have lagged in their rhetoric. In the last 20 years, the public rhetoric surrounding climate change has changed from one based on mainly economic arguments to ecological ones (Ylä-Anttila et al, 2018). Increasingly, climate change is seen not as hindering economic growth, but as a prerequisite for it to be sustained in the future. Climate change can be seen as something that regulates more and more human activities, and that makes new acts of collective intentionality possible — it adds functions to it. If these functions begin to affect general policy, they acquire a normative status and become constitutive rules (Searle, 1995, p. 48). As we have an increasing amount of policies addressing climate change, the phenomenon is becoming increasingly normative. The term is utilized as a buzzword involving economic opportunities promoting growth, and at the same time it has been connected to movements aiming to limit the amount of freedom individuals have as consumers, such as flight tax proposals and minimalist lifestyles. These new forms of collective actions help enforce climate change as a social object.

As we form a clearer global consensus on the idea that something has to be done to tackle climate change, this allows us to propose regulative and constitutive rules to guide an increasing amount of actions. These include the above-mentioned tax propositions, but also larger, international agreements on methods to curb global warming and increased $CO_2$ emissions. These in part direct an ever increasing part of humans works to be involved in areas related to fighting climate change, such as in being employed as a manufacturer of air turbine parts, or as analysts aiming to optimize heavy traffic to minimize transport emissions, again further enforcing the social object in question.

## 2. Climate change governance and deliberative democracy — A.M. Taskinen

Climate change is a complex phenomenon, which causes humankind many problems when trying to govern it. The major problem concerning climate change governance is that it affects many countries and many people around the world. How can everyone have their voices and viewpoints heard? Some researchers think that the best way to treat this challenge is to use the tools of deliberative democracy (e.g.

Niemeyer, 2013). In this section, I will discuss the definition of deliberative democracy, consider the challenges and possibilities that deliberative democracy has in climate change governance and discuss climate change governance and deliberative democracy in local and international arenas.

### 2.1 What is deliberative democracy?

In Chapter 1 of *The Oxford Handbook of Deliberative Democracy* (2018) authors Bächtiger, Dryzek, Mansbridge, and Warren discuss the concept of deliberative democracy. At the heart of the concept lies deliberation: public discussion that helps people to follow public problems better (Bächtiger, Dryzek, Mansbridge & Warren, 2018, 2). They also point out that the concept of deliberative democracy is aspirational, which means that it is more of an ideal to aim for (Bächtiger et al., 2018, 2). They describe how the ideals of deliberative democracy have changed during the years between first-generation and second-generation thinkers (Bächtiger et al., 2018, 3). The main differences between the two generations' ideals concern equality, reasons, consensus and common good: second-generation thinkers underline "inclusion, mutual respect," "relevant considerations", "clarifying conflict" and "orientation to both common good and self-interest constrained by fairness" more than first-generation thinkers (Bächtiger et al., 2018, 4, table 1.1). According to Bächtiger et al. (2018, 3), the second-generation ideals fit modern 21st-century deliberation better. Niemeyer (2013, 430) also points out that deliberative democracy must also be consequential, which means that deliberation must affect the final decisions.

### 2.2 Climate change governance and deliberative democracy: possibilities and challenges

A big problem with climate change governance is that climate change is easy to ignore in discussions because it can feel too complex of a phenomenon to understand (Niemeyer, 2013, 431). Climate change as a phenomenon can appear very blurred to many, so concrete actions to tackle it may even create resistance (Niemeyer, 2013, 432). Niemeyer (2013, 433) points out that especially for regular citizens it may be hard to understand a vast phenomenon like climate change and demand changes from the political elites. And even if the demands were to be expressed, it can be hard for democracies to convert people's demands to concrete actions (Niemeyer, 2013, 433).

However, Niemeyer (2013) states that the best way to meet these challenges of climate change governance is deliberative democracy. According to him, these challenges can be solved by using the tools of deliberative democracy in public discourse (Niemeyer, 2013, 431). Public deliberation can help to articulate environmental issues and make their importance salient (Niemeyer, 2013, 434). Besides, Niemeyer (2013, 432, 433) also says that the democratization of climate change discussion pays off because political elites are not as interested in the environment as people in civil society**.**

Niemeyer (2013, 435) states that the main challenge for deliberative democracy is to have a proper discussion environment where deliberation can achieve something: the solution for Niemeyer is mini-publics. The study from Australian mini-public shows that deliberation affected what participants thought about climate change as well as the way they discussed climate change (Niemeyer, 2013, 441). The discourse changes show that participants had more consensus (still with a hint of diversity) and more adaptive opinions after deliberation (Niemeyer, 2013, 442; Hobson and Niemeyer 2011, 966).

According to Niemeyer (2013, 442), the main benefit of deliberation in the study was that it helped participants have a clearer picture of the complex issues related to climate change. Deliberation also changed the way participants viewed climate change governance: after deliberation, they thought that citizens can govern climate change in cooperation with the government, making it feel like a more collective problem, not just a problem of political leaders (Niemeyer, 2013, 443, 448; Hobson & Niemeyer, 2011, 968). Although the study from Australian mini-public had some good effects on the way participants viewed climate change governance, Niemeyer (2013, 444) also says that it might be challenging to create the same, working setting for deliberation in other, larger settings.

So, Niemeyer's (2013) discussion shows that even if deliberation in mini-publics could be helpful in using deliberative democracy's tools in climate change governance, the problem of adding them to larger, even international settings, remains. Next, I will discuss how well climate change governance fills the ideals of deliberative democracy at the international level and also take a look at deliberative democracy's benefits and challenges at local levels.

### 2.3 Climate change governance and deliberative democracy: local and international arenas

Bächtiger et al. (2018, 9) also point out that deliberative democracy can take place in various places. Deliberative democracy's arenas can be local or international, formal or informal, and may take place in governmental institutions or civil society (Bächtiger et al., 2018, 10-13).

Dryzek and Stevenson (2011, 1867, 1868) evaluate the global governance of climate change based on the deliberative system. It includes the following parts: "the public space", "empowered space", "transmission", "accountability", "meta-deliberation" and "decisiveness" (Dryzek & Stevenson, 2011, 1867, 1868). According to Dryzek and Stevenson (2011, 1873), there are various discourses represented in public space of international arenas, which is a good thing. But as for empowered space, there is only little deliberation and the transmission of civil society's input to governmental bodies does not work as well as it could because some discourses are neglected (Dryzek & Stevenson, 2011, 1873). Also, accountability, meta-deliberation, and decisiveness of the system do not work very well (Dryzek & Stevenson, 2011, 1873). In another study, there are implications of non-governmental actors bringing alternative views to international arenas, but also observations that despite their participation, some views still get marginalized (Nasiritousi, Hjerpe & Buhr, 2014, 183). So, Dryzek and Stevenson's (2011) and Nasiritousi, Hjerpe and Buhr's (2014) studies seem to underline the problem that Niemeyer (2013, 444) formulates: good deliberation is hard to achieve in large, real-life settings.

Above I discussed Niemeyer's (2013) arguments of the benefits of mini-publics. Similarly, Romsdahl and Kirilenko (2018, 278) think that it is necessary to use deliberative tools in local governance levels, as well. According to Romsdahl and Kirilenko (2018, 284) benefits of deliberation can emerge, for example, in countries where there is climate change skepticism in governmental bodies: they think that through deliberation in civil society there is a possibility to come up with solutions to climate change problems. In some countries, there can also be a possibility for fruitful cooperation between the national government and different local actors (Romsdahl & Kirilenko, 2018, 284). Local deliberation can help to articulate regional climate problems (Romsdahl & Kirilenko, 2018, 284). Romsdahl and Kirilenko (2018, 284) also point out the problems regarding local deliberation: among other things, there can be only little interest to participate, inequality in participating and even resistance from different directions.

### 2.4 Conclusion

At the end of this section, I would like to reiterate the point made by Bächtiger and others (2018, 2): the ideals of deliberative democracy are primarily aspirational, and they can be hard to fulfill in real-life settings. But this does not mean that we should lose hope concerning deliberative democracy in climate change governance. All of the researchers, whose papers I have examined in this section, seem to agree on some level that deliberative democracy is worth chasing in climate change governance, both at the local and international levels. Of course, they have justifiably pointed out various problems and challenges of deliberative democracy in climate change governance — but it is certainly worth aiming for better practices in climate change governance and trying to achieve the ideals of deliberative democracy as well as possible.

## 3 Transnational climate change governance: why won't we cooperate? Perspectives from game theory — R.M. Savo

Climate change is a transnational problem which needs the cooperation of multiple countries, making the problem difficult to solve. All countries want to hold on to their autonomy and decide for themselves, and attempts to create a global plan to stop climate change have not been successful in the past. There is a need for a theory that would help countries plan strategic actions to cut down emissions globally. One useful perspective in solving this cooperation issue is a theory that is often used in social and political sciences: game theory.

### 3.1 Prisoner's dilemma and the collective good

Game theory helps to explain strategic behavior of rational individuals with the help of mathematics (Wood, 2011, 153). It allows us to understand the clash of individual and collective action in global governance of climate change. Researchers have argued that the reason why it is so hard to find a solution for climate change is that there is a strong incentive to "free-ride", meaning that one country will not cut down their pollution rates while others do (Wood, 2011, 153; DeCanio & Fremstad, 2013, 180). According to game theory, this problem arises from the fact that what is best for an individual country is not the same as what is best for all the countries collectively (Wood, 2011, 153). This idea was first put forward by Garret Hardin in 1968, and it has been known as "the tragedy of the commons" in environmental science: in a world of shared resources individuals act on their self-interest and end up spoiling the common good.

A classic example of game theory is the Prisoner's dilemma. There are two people who have committed a crime and they are being interviewed separately. Their sentence depends on what information they share. If they both say they didn't do it, they both serve 1 year. If they say that the other person did it, they can avoid prison but the other person will get a 10 year sentence. However, if they both blame the other person, they will both go to jail for 5 years.

The issue in Prisoner's dilemma is that what is best for the individual is not the same as what is best collectively. The best option would be for both of them to stay silent and collectively serve 2 years. But no matter what each person will do, it is best for both individuals to say that the other person committed the crime (Amadae, 2016, 28). Game theorists often come to the conclusion that in the end everyone will tell on the other person, because it is the only way to be sure that you get a good outcome (Amadae, 2016, 29; Wood, 2011, 155; DeCanio & Fremstand, 2013, 180). This solution, where it's not possible for one actor to get a better end result by changing their own behavior, is called the Nash equilibrium (DeCanio & Fremstand, 2013, 179). However, many times the best solution needs cooperation. Cooperation, on the other hand, needs communication and an agreement between the parties involved (Wood, 2011, 156-157).

The Prisoner's dilemma is a classic example of two people making decisions that affect both of them, but it has also been used to understand political actions and the world (Amadae, 2016, 25). With climate change all the countries would be better off if they cooperated and created a plan that everyone promised to follow (Amadae, 2016, 231). However, there are multiple reasons why that is not happening.

First, there is a constant competition over economic and political power, and if one country starts cutting down pollution, it becomes less productive than others (DeCanio & Framsted, 2013, 182). Thus, as long as there is no binding agreement, it is better for an individual country to keep polluting, as in the Prisoner's dilemma (Wood, 2011, 155). Second, global politics is much more complicated than the Prisoner's dilemma: it involves different cultures and values as well as complex power structures. DeCanio and Fremstand (2013, 183) speak of "different moral universes", pointing out that developing countries and developed countries are in very different positions when it comes to climate change.

Third, there are "superpowers", such as the United States and China, which have much more influence over climate change, but which are more interested in keeping their power than saving the planet. It's possible that the heads of leading countries have not understood or accepted the seriousness of climate change (DeCanio and Fremstand, 2013, 182). In the end there is the same Nash equilibrium as with the Prisoner's dilemma: what is best for an individual country is always to keep polluting, because no matter what other the other countries are doing, everyone wants to win (Amadae, 2016, 231).

### 3.2 How can we increase cooperation?

Wood (2011, 160) suggests multiple ways to increase cooperation between actors in this environmental dilemma. Countries would be more likely to cooperate if there were a binding contract, as long as those countries that would not obey would get punishments. Because economics are of high priority, there could be trade and taxation contracts to give benefits to those countries that follow the rules. In addition, if the situation is repeated, it is more likely that actors will cooperate. The problem with climate change is that on one hand it is a repeated game, because it is repeatedly negotiated. On the other hand, it is not a repeated game but an ongoing process of increasing emissions, which affect future generations (Wood, 2011, 160).

Amadae (2016) cites the *Stern Review*, which has suggested steps to increase cooperation in collective action. She suggests that incentives should be changed so that committing to an agreement would result in something positive. For example, the country's reputation would improve as a result of joining. In addition, there should be more repeated interactions between parties to increase trust, and those that don't cooperate should be punished. Lastly, there should be a possibility to renegotiate agreements. (The *Stern Review*, here Amadae, 2016, 227)

DeCanio and Fremstad (2013) try to look at the issues from another point of view: how can we implement a strategy that would lead to the Nash equilibrium when all countries cut down pollution? They argue that the Prisoner's dilemma is not the best way to look at the climate debate. Battling climate change should be a *Coordination Game* where the best end result for all the parties is when they cooperate. If climate change is seen as a threat to the survival of humankind, it is undeniable that working together to save the planet is the best option. However, if it all comes down to winning the geopolitical race, modeled as Prisoner's dilemma, it's always best to keep on polluting (DeCanio and Fremstad, 2013, 182).

What seems to be the underlying issues preventing transnational agreement are the ignorance of the seriousness of climate change, the varying "moral universes" as DeCanio and Fremstad argue, as well as the constant battle to "win" the economic and political global game. Game theory offers actions to take in order to increase cooperation between countries, most importantly making it beneficial to join an international climate agreement and adding the amount of communication between the parties involved. Because countries are most interested in maintaining economic power, economic incentives, such as taxation towards polluting countries, would make an agreement seem more attractive to countries.

## 4 The approach to advance the collective intelligibility of climate change: the possibility of melding two discourses in news reports — D. Huo

Journalism has been a source for information for hundreds of years. As an important institution, it could add vital dimensions to discussions about the complex phenomenon of climate change.

### 4.1 Journalism as a main vessel of disseminating information

Agenda setting is a theory introduced by Maxwell McCombs and Donald Shaw in a seminal study conducted during the 1968 elections in the United States. At first, it showed how editors, newsroom staff, and broadcasters influence public opinion and shape political reality (McCombs & Shaw, 1972). Nowadays it has expanded to include several other aspects beyond the transfer of salience of issues from the media agenda to the public community (Valenzuela, 2019).

In recent decades, mass media, responsible as they are for information in all areas, have taken it upon themselves to provide information with ever increasing intensity on scientific progress. The social representation of scientific knowledge is actually derived in large part from the news media, which can be seen in climate change reporting. At the heart of climate change is the proposition that human activities are altering the composition of the planet's atmosphere to a degree sufficient to affect the natural processes that play fundamental roles in shaping the global climate (Trumbo, 1996, 270). And climate change has gone from a vague environmental concern several decades ago to a confirmed global phenomenon that is today affecting virtually every aspect of our society — our economy, security, health, livelihoods, food supply and our politics — and as such it has become ever more ripe for investigation (Fahn, 2019). Since the public garners knowledge about science from the mass media (cf. Nelkin & Elias, 1996; Wilson, 1995), investigating the portrayal of climate change resulting in collective intelligibility is crucial. As a main resource for the public, journalism plays a vital role in disseminating information about climate change, attracting increasingly more attention from the public and developing collective intelligibility of the complex challenge it poses.

### 4.2 The gap between popular discourse and scientific discourse

A discourse can be understood as a set of categories and concepts, which enables the mind to process sensory inputs into coherent accounts, which can then be shared in intersubjectively meaningful fashion (Dryzek & Niemeyer, 2008, 481). Van Dijk (1993) argues that there is some relevance of a socio-cognitive interface between discourse and dominance, which is based on underlying historical, social, political and cultural properties. According to agenda setting theory, it seems easy to conclude that the more climate change is mentioned in reports, the more the public would pay attention to it, but whether the reports are working as they are supposed to is open to query. As there is a gap between science and society, the communication of scientific knowledge is different to the general public. In comparison to popular discourse, scientific discourse is an encoded form of knowledge that requires translation in order to be understood (Ungar, 2000, 298). Some people believe that with the dissemination of scientific articles, the gap between the realistic situation and public knowledge would narrow, and finally help to deal with climate change (Schoenfeld, Meier & Griffin, 1979). This shows journalism's mission, exploring facets of the discourse characteristically used to present scientific knowledge to the general public and explaining opaque concepts and theories to us.

Some former research from four corners of the globe also demonstrates the problem (cf. Boykoff & Boykoff, 2004; Takahashi, 2011; Dunwoody & Peters, 2016; Chen, Ghosh, Liu & Zhao, 2019). This research shows that the coverage of global warming has contributed to a significant divergence of popular discourse from scientific discourse. Reichel (2018) is a journalist who also listed some issues in today's coverage of climate change. For instance, reporters may focus on subjects that make it hard for readers to relate. It is no doubt that such journalistic norms would lead to negative outcomes. Such reports might not work for improving the audience's understanding. But further, without specifically constitutive rules, for instance, that the core of climate change reports is to be understood by the majority of the audience, a lack of basic collective intelligibility could result, thus evoking insufficient consensual coordination of social practices.

### 4.3 How could news coverage help to govern climate change?

When we are talking about climate change, the core idea is that we should have a picture of what it is, so the first mission of the news is to tell the audience what is happening with our planet. In other words, journalism should help the public by encoding information. And it is obvious that the coverage should take the responsibility to promote the formation of collective intelligibility by bridging popular and scientific discourse.

For melding such two discourses, journalism might need to review their professional norms and consider changes happening in the economy, technology and our society. It would be necessary to talk about the relationship between journalistic norms and communications strategy. There is no doubt that professional norms play a vital role. Basing the news on such rules, we could get fair and informative coverage. But qualified reports need to be understood by the public, especially for climate change coverage, and journalism should adjust the relationship between meaning and strategies.

Paying attention to audience-centered communication might be a practical method, with its emphasis on “the meeting place between scientific discourse and common readers” (cf, Allen, 1992). Firstly, the translation of climate change into concrete and relevant terms of daily life is an operational attempt. This is achieved with various patterns via a connection to the recipients’ everyday experiences and perceptions (Weingart et al., 2000). For instance, editors can expand climate change coverage beyond the science desk. As an example, the energy consumed in transportation is a great daily angle to introduce climate change and simultaneously provide simple tips for readers. Moreover, finding individual characters to tell their stories is also practical. Asking employees and professors to tell their stories might help the audience to understand the message. Secondly, journalism could take advantage of technology. The new scenario of the digital technologies could reconsider on a new basis the theme of a complex real structure. Visual presentation might open up avenues for scientific news, especially in our current digital era. According to the research done by Usher (2016), journalism shows the trend of using online databases, and data interactives and visualizations, which might lead to possibilities of investigating the way to improving understanding of scientific discourse. Moreover, the development is actually a novel form of interactive storytelling (Royal, 2012, 20; Usher, 2016), which might create a bridge between the public and science. Automated journalism might also help, fitting different people’s education, occupations and other social conditions to help the public understand climate change better.

Overall, journalism is supposed to find a balance between the obscure scientific articles and public awareness, which is essential to achieve collective intelligibility and to facilitate climate change governance. Journalism could help audiences become aware of the relevance of climate change to their own behavior, as well as of the immediacy of climate change as a global environmental problem. Media coverage should lead to more understanding and discussion, and facilitate climate protection efforts.

# 5. Algorithmic governance — A solution to the climate crisis? — R.M. Kontiokari

As previous chapters have shown, climate change should be understood as a number of complex physical and social phenomena and hence it is challenging to govern. Excluding a few loosely binding contracts, such as the Kyoto Protocol and the Paris Agreement, there is still a lack of a transnational authority that would be responsible for governing climate issues and coordinating global climate action (e.g., Bulkeley 2016).

There is an urgent need for global operations to reduce emissions and mitigate the already-present adverse effects of global warming; however, the structures upholding the unsustainable system are so deeply integrated into the global political economy, that even the establishment of large-scale measures against climate change has been proven to be extremely difficult (e.g. Stevenson & Dryzek, 2014).

Artificial intelligence (AI) and machine learning (ML) have proven to be useful tools for mitigating pollution and adverse climate effects on a local scale (e.g., Rolnick et al., 2019) However, involving AI and algorithmic decision-making in broader climate *governance* to create comprehensive guidelines and solutions for global climate action is still a relatively new concept. Therefore, in this chapter, my aim is to explore the potential that *algorithmic governance,* or *algocracy,* has for the fight against climate change. First, I will begin with a brief overview of the current state of climate governance and its main problems; second, continue to explain the concept of algorithmic governance; third, explore the idea of algorithmic decision-making as the basis for global climate governance; and last, finish with a short conclusion.

## 5.1 Global climate governance and the authoritarian temptation

Although there has been scientific proof of anthropogenic climate change since the late 19th century, there is little to no global governance of the issue (Stevenson & Dryzek, 2014, 2). Much of the problems in global climate governance are rooted in the current global system stressing national sovereignty, which seems to be "poorly equipped" to match the urgency and magnitude that is required for sustainable and effective climate action (Stevenson & Dryzek, 2014, 4). Nation-states seem to be incapable of reaching the consensus on what should be done; who should we listen to; who is responsible; and on top of it all, whether the changes in the climate are even caused by human action (Turnheim, Kivimaa & Berkhout, 2018).

Due to this hindrance, there have been calls for authoritative climate governance that would overrule national democracies and their sovereignty, parallel to Lovelock's *Gaia Hypothesis* (Stevenson & Dryzek, 5). However, the issue with such initiatives is that there is no global body or democracy that could be overruled or "put on hold"; in other words, global climate governance cannot be realized, as it yet needs to be established (Stevenson & Dryzek, 5).

Another issue with the idea of authoritarian climate government is that it perceives climate change as a problem that needs to be solved. However, considering how deeply anthropogenic climate change is, integrated with societies and human action, it should be seen as a condition that inevitably comes with "our sociomaterial order and domain" (Bulkeley 2016, 167). Therefore, a more sustainable approach to governing climate change would be to perceive climate governance as an "ever-expanding activity" that increasingly involves more and more diverse elements of humanity and society, aiming collectively towards more sustainable societies (Bulkeley, 2016, 155). As stated in chapter one of this paper, such activities do already exist on a smaller scale; we have climate strikes, recycling initiatives, different environmental taxes, and education on less energy-intensive ways of living. What is lacking is a global climate authority that would set the rules and regulations that different actors should and would be willing to follow.

### 5.2 Algorithmic governance

Recent rapid changes in our technological and communicative environments have increased the amount of computable data significantly, and consequently created possibilities to aggregate knowledge and solutions through different AI applications. Algorithms are capable of processing massive amounts of data and producing solutions much faster than humans. Thus, there has been "growing willingness" to include algorithm-based decision-making systems into democratic governance of societies (Danaher et al., 2017, 2).

Governance, by itself, refers to institutional steering that includes "horizontal and vertical extension of traditional government" (Latzer & Festic, 2019); in other words, a combination of all the processes of governing. Governance is not limited only to rules and regulations by national governments but can also be understood as rules and norms in different social systems.

Algorithmic governance, in turn, refers to both the involvement of algorithmic decision-making processes in governance, and the "intentional and unintentional steering effects" that "uphold selection systems" in our technological arenas (Latzer & Festic, 2019). The term also includes the "intentional attempts to manage risk or alter behavior in order to achieve pre-specified goals" (Yeung, 2018, 3). Therefore, algorithmic governance should also be separated from a total authoritarian *algocracy*, a society that would be fully governed by some mysterious "master algorithm" that is autonomous from human action and capable of making just decisions for all of humankind (Danaher et al., 2017). In fact, it would be foolish to assume that algorithms would be capable of such autonomy, as it is nearly impossible to create an algorithm free from human impact in a human environment. As long as algorithms are systems constructed by humans, they will carry biases from the environments they were created in. There are also issues with the opacity of algorithms, their reliability, and compatibility with current governing systems, that makes the idea of an autonomous master algorithm rather undesirable (Danaher et al., 2017).

### 5.3 Algorithms as guidelines for climate governance

One of the primary issues with involving algorithms in climate governance is overcoming the "risk of meaningless machines" (Latzer & Festic, 2019). An algorithm can surely produce a great number of solutions to how humankind could reduce carbon emissions and decrease overall pollution, but how can we be sure that such solutions would be socially sustainable, and executable in realistic socio-technological contexts? For instance, an essay was recently published in *The Economist* on tackling climate change that was written by an algorithm (*The Economist*, 2019). It was appreciated for its clarity and reliability, but criticized for the vagueness of its solutions, such as "alternative economy". The solutions produced by the algorithm did not differ much from solutions made by climate experts, and thus this experiment hints that AI may not be capable of solving climate change on its own or serving as a guiding, governing authority on a global scale.

However, if we comprehend climate change as a permanent condition and its governance as a number of ever-expanding activities (Bulkeley, 2016), AI might have its benefits. As there is proof that AI can successfully help in reducing pollution or mitigating its adverse effects locally (Rolnick et al., 2019, 2), it is not far-fetched to suggest that a number of AI solutions could be helpful in the construction of global climate governance. For instance, AI could be used to create a roadmap on how different human actions actually affect the climate and what kind of effects the changing climate has on current human activities. This group of solutions would ultimately form a guideline representing climate realities that societies could refer to in deliberative democracies. Therefore, these AI implementations should not only exist on a global scale, but also vertically at different levels of regulation.

The tentative suggestion above, however, is yet superficial and flawed. Some considerations that should be made are whether the mentioned set of AI solutions would actually be mutually compatible and comprehensible; if the suggestions produced on different levels are contradictory, then AI does not bring much value into climate governance. In addition, the biases and power structures related to the creation and implementation of AI should be taken into serious consideration.

### 5.4 Conclusion

In conclusion, there are many benefits of implementing algorithmic decision-making into global climate governance. Global climate governance would be most effective as a system that considers climate change as an all-encompassing condition that needs constant assessment in all domains and practices of societies. However, many practical issues remain around implementing algorithms, such as accountability, accessibility, and property rights issues.

## Conclusion

In this research paper we considered the challenges and possibilities of governing climate change from various viewpoints. Climate change has become a social phenomenon, which demands all humans and countries to take a stance. The big challenge with governing climate change is the vast pool of complex information circulating and the amount of different actors involved. There is a need for a shared understanding of the social reality of climate change: the economic discourse needs to move to an idea about a common world that we have to protect.

Game theory points out that finding a common path with climate change is difficult, because there is a great temptation to "free ride". Because individual countries look for their own self-interest, we have to create a climate agreement that is attractive, and the best option for all countries. Countries will not join any agreement "just to save the world". Governments have a big responsibility in governing, but citizens can affect governance by means of deliberative democracy. When citizens take part in public discussions and cooperate with governments, there is hope for a better understanding of the issues at hand and action to be taken.

Publics gets much of their knowledge of climate change from different media, and journalism is a central actor in translating and conveying complex scientific information to the public. By expanding climate change coverage outside the science desk and using new technologies such as interactive storytelling, journalism could enhance the discussion around climate change. One future hope is the use of algorithmic governance to help humans cope with the masses of information and actions. With the help of algorithms, we could create a more comprehensive plan for climate change — one that encompasses all human activities and creates a more sustainable way of life across nations.

## References

Allan, B. (2017). Second Only to Nuclear War: Science and the Making of Existential Threat in Global Climate Governance. *International Studies Quarterly, 61*(4). Retrieved from https://academic.oup.com/isq/article-abstract/61/4/809/4769244?redirectedFrom=fulltext

Allen, R. (1992). *Channels of discourse, reassembled: Television and contemporary* (2nd ed.). Chapel Hill: University of North Carolina Press. 101–103

Amadae, S. M. (2016). *Prisoners of Reason: Game Theory and Neoliberal Political Economy*. New York: Cambridge University Press.

Anderson, B. (1991). Imagined communities reflections on the origin and spread of nationalism. London: Verso.

Bächtiger, A., Dryzek, J. S., Mansbridge, J. & Warren, M. E. (2018). Chapter 1: Deliberative Democracy, An Introduction. In Bächtiger, A., Dryzek, J. S., Mansbridge, J. & Warren, M. E. *The Oxford Handbook of Deliberative Democracy*, pp. 1–27, Oxford: Oxford University Press.

Boykoff, M. T., & Boykoff, J. M. (2004). Balance as bias: global warming and the US prestige press. *Global Environmental Change, 14*(2)*,* 125–136. doi:10.1016/j.gloenvcha.2003.10.001

Bulkeley, H. (2016). *Accomplishing climate governance.* Cambridge: Cambridge University Press.

Chen, Y., Ghosh, M., Liu, Y., & Zhao, L. (2019). Media Coverage of Climate Change and Sustainable Product Consumption: Evidence from Hybrid Vehicle Market. *Journal of Marketing Research*, *56*(6), 995–1011. doi:10.1177/0022243719865898

D'Aprile, A. (2018). Not Enough: Latest Studies put Pressure On COP24 To Speed-Up Climate Action. Retrieved from: https://www.climateforesight.eu/global-policy
/not-enough-latest-studies-put-pressure-on-cop24-to-speed-up-climate-action/

Danaher, J., Hogan, M. J., Noone, C., Kennedy, R., Behan, A., De Paor, A., … Shankar, K. (2017). Algorithmic governance: Developing a research agenda through the power of collective intelligence. *Big Data & Society, 4*(2), 1–21. https://doi.org/10.1177/20
53951717726554

Decanio, S. J. & Fremstad, A. (2013). Game theory and climate diplomacy. *Ecological Economics, 85*, 177–187. doi:10.1016/j.ecolecon.2011.04.016

Downs, A. (1973). Up and down with ecology. *The "issue attention" cycle.*

Dryzek, J. S. & Stevenson, H. (2011). Global democracy and earth system governance. *Ecological Economics, 70*(11), pp. 1865–1874. doi:10.1016/j.ecolecon.2011.01.021

Dryzek, J. S., & Niemeyer, S. (2008). Discursive representation. *American political science review*, *102*(4), 481–493. doi:10.1017/s0003055408080325

Dunwoody, S., & Peters, H. P. (2016). Mass media coverage of technological and environmental risks: A survey of research in the United States and Germany. *Public understanding of science*, *1*(2), 199–230. doi:10.1088/0963-6625/1/2/004

Fahn, J. (2019). Retrieved October 30, 2019 from https://gijn.org/2019/04/22/investigating-the-story-of-the-century/

Griswold, E. (2012). How 'Silent Spring' Ignited the Environmental Movement. Retrieved from https://www.nytimes.com/2012/09/23/magazine/how-silent-spring-ignited-

Hobson, K. & Niemeyer, S. (2011). Public responses to climate change: The role of deliberation in building capacity for adaptive action. *Global Environmental Change*, *21*(3), pp. 957–971. doi:10.1016/j.gloenvcha.2011.05.001

IPCC. (2018). Global Warming of 1.5°C. An IPCC Special Report on the impacts of global warming of 1.5°C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty. Retrieved from https://www.ipcc.ch/sr15/

Latzer, M. & Festic, N. (2019). A guideline for understanding and measuring algorithmic governance in everyday life. *Internet Policy Review, 8*(2). doi: 10.14763/2019.2.1415

McCombs, M. E., & Shaw, D. L. (1972). The agenda-setting function of mass media. *Public opinion quarterly*, *36*(2), 176–187. doi:10.1086/267990

Nasiritousi, N., Hjerpe M. & Buhr K. (2014). Pluralising climate change solutions? Views held and voiced by participants at the international climate change negotiations. *Ecological Economics, 105*(C), 177–184. doi:10.1016/j.ecolecon.2014.06.002

Nelkin, D., & Elias, J. (1996). Selling Science: How the Press Covers Science and Technology (revised edition). *Journal of Public Health Policy*, *17*(4), 501–503.

Niemeyer, S. (2013). Democracy and Climate Change: What Can Deliberative Democracy Contribute? *Australian Journal of Politics & History, 59*(3), 429–448. doi:10.1111/ajph.12025

Reichel, C. (2018, May). Covering climate change: What reporters get wrong and how to get it right. Retrieved October 31, 2019, from https://journalistsresource .org/tip-sheets/reporting/climate-change-reporting-tipsheet-elizabeth-arnold/

Rolnick et al. (2019). *Tackling Climate Change with Machine Learning.* Cornell University.

Romsdahl, R., Blue, G. & Kirilenko A. (2018). Action on climate change requires deliberative framing at local governance level. *Climatic Change*, *149*(3–4), 277–287. doi:10.1007/s10584-018-2240-

Royal, C. (2012). The journalist as programmer: A case study of the New York Times interactive news technology department. *The Official Research Journal of the International Symposium for Online Journalism, 2*(1), 5–24.

Scheufele, D. A., & Tewksbury, D. (2006). Framing, Agenda Setting, and Priming: The Evolution of Three Media Effects Models. *Journal of Communication, 57*(1), 9–20.

Schoenfeld, A. C., Meier, R. F., & Griffin, R. J. (1979). Constructing a social problem: the press and the environment. *Social problems*, *27*(1), 38–61. doi:10.2307/800015

Stevenson, H. & Dryzek, J. (2014). *Democratizing Global Climate Governance.* Cambridge University Press.

Stripple, J. & Carlsson, S. (2003). Climate Governance Beyond the State. *Global Governance*, 9(3), pp. 385–399. doi:10.1163/19426720-00903009

Takahashi, B. (2011). Framing and sources: A study of mass media coverage of climate change in Peru during the V ALCUE. *Public Understanding of Science*, *20*(4), 543–557. doi:10.1177/0963662509356502

The Economist. (2019). *How to respond to climate change, if you are an algorithm.* Retrieved from https://www.economist.com/open-future/2019/10/01/how-to-respond-to-climate-change-if-you-are-an-algorithm

The Intergovernmental Panel on Climate Change. (2018). Global Warming of 1.5 C. Retrieved from https://www.ipcc.ch/sr15/

The White House. (1965). Restoring the Quality of Our Environment. Retrieved from: https://ozonedepletiontheory.info/Papers/Revelle1965AtmosphericCarbonDioxide.pdf
the-environmental-movement.html

Trumbo, C. (1996). Constructing climate change: claims and frames in US news coverage of an environmental issue. *Public understanding of science*, *5*, 269–283. doi:10.1088/0963-6625/5/3/006

Turnheim, B., Kivimaa, P., & Berkhout, F. (Eds.). (2018). *Innovating Climate Governance: Moving Beyond Experiments.* Cambridge University Press. doi: 10.1017/9781108277679

Ungar, S. (2000). Knowledge, ignorance and the popular culture: climate change versus the ozone hole. *Public Understanding of Science*, *9*(3), 297–312. doi:10.1088/0963-6625/9/3/306

Usher, N. (2016). Interactives Journalism: Hackers, Data, and Code. Champaign: University of Illinois Press.

Valenzuela, S., (2019). Agenda Setting and Journalism. Oxford Research Encyclopedia of Communication. Retrieved 30 Oct. 2019, from https://oxfordre.com/com munication/view/10.1093/acrefore/9780190228613.001.0001/acrefore-9780190228613-e-777.

Van Dijk, T. A. (1993). Principles of Critical Discourse Analysis. *Discourse & Society, 4*(2), 249–283. doi:10.1177/0957926593004002006

Weingart, P., Engels, A., & Pansegrau, P. (2000). Risks of communication: discourses on climate change in science, politics, and the mass media. *Public understanding of science*, *9*(3), 261–284. doi:10.1088/0963-6625/9/3/304

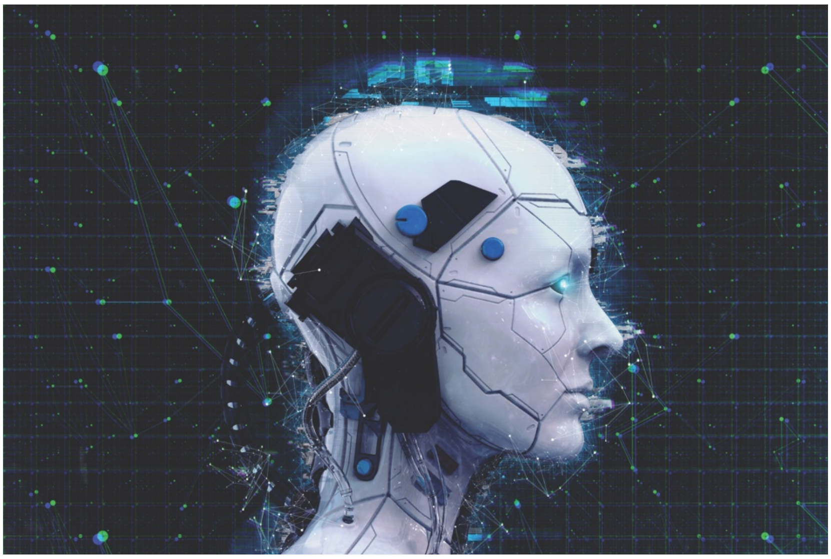
Wood, P. J. (2011). Climate change and game theory. *Annals of the New York Academy of Sciences, 1219*(1), 153. doi:10.1111/j.1749-6632.2010.05891.x

Yeung, K. (2018), Algorithmic regulation: A critical interrogation. *Regulation & Governance,* 12, 505–523. doi: 10.1111/rego.12158

Ylä-Anttila, T., Vesa, J., Eranti, V., Kukkonen, A., Lehtimäki, T., Lonkila, M., & Luhtakallio, E. (2018). Up with ecology, down with economy? The consolidation of the idea of climate change mitigation in the global public sphere. *European Journal of Communication*, *33*(6), 587–603. https://doi.org/10.1177/0267323118790155

# Part IV

# Digital Emancipation



# 4.1 How Digital Innovations of Mobility Can Emancipate Women in the Global South — Case Study on Ride-Hailing Apps

Vilma E. Lappalainen, Lasse I. Keinonen, Riikka T. Ilmonen, Tiina M. Helojärvi,
Saana H. Annala
Faculty of Social Sciences, University of Helsinki

## Abstract

Technological development is often associated with a possibility of emancipation and freedom for individuals. How can new advancements of mobility, such as ride-hailing technologies in the form of mobile applications liberate women in developing countries? In this research project, we dive deep into the emancipatory effect of Uber and other similar applications. We are introducing two case studies: one from the perspective of female drivers, and another one from the passengers of ride-hailing apps. We are going to approach this discussion from a perspective that combines feminist technoscience and viewpoints of value-sensitive design method. We conclude with policy recommendations, where we evaluate our findings according to the Principles of Digital Development. We find that ride-hailing apps have some emancipatory effects, but more research is needed before further conclusions. We suggest that ethically sustainable ride-hailing services in the Global South must study the users and the ecosystem and build on already existing technologies.

*Keywords*: Feminist technoscience; technology; mobility; emancipation; ride-hailing apps, Uber, Lyft, STS, science and technology studies

In the 21st century, mobility has become more and more accessible as moving from one place to another is now possible by using smartphone applications. Applications such as Uber and Lyft have quickly been able to establish their place as everyday tools for many people. Even with their positive effect on mobility, these kinds of applications have gained some negative press in recent years (e.g. *The Guardian*, 19 June 2019). For this research project, we wanted to focus on what emancipatory effects ride-hailing applications might have for women in countries where their mobility has traditionally been more limited.

Our main focus here are women in the Global South[1]. However, Western data is also partially used because some of the data can only be found in an occidental context. From the perspective of female drivers on ride-hailing apps, we are focusing on women as drivers in general, and for female passengers, we are focusing on women travelling in India, Mexico and South Africa. We consider this somewhat ambivalent use of context and data to be justified.  This is because our focus is on the political design problems related to the ride-hailing apps (which can be context bound) and to a lesser degree on research about the Global South. Focus on the Global South is used specifically to highlight context and North-South power differential and related challenges.

We assume that ride-hailing apps have had manifold and contradicting effects for female emancipation. This is because during the design process, the developers have not been able to imagine all the possible problems related to the use of their products — the developers might not even be interested in taking these into account. Some of the shortcomings may stem from the fact that the apps were originally developed to serve a particular purpose in a particular environment, but through temporal and spatial change, new challenges have arisen.

In the following section Keinonen establishes our theoretical framework which combines feminist technoscience and value-sensitive design approach (VSD). In the third section Lappalainen discusses the relationship between mobility and female emancipation. Then we move on to our case studies from two different perspectives: Annala analyzes ride-hailing apps from the driver's perspective and in turn Ilmonen takes the angle of the passenger. In the sixth section Helojärvi uses the results of the previous sections to make policy recommendations according to the framework of Value-Sensitive Design and the Principles of Digital Development[2]. We end our research paper with conclusions.

[1] We are using both the terms Global South and developing countries, but we find the term Global South more appropriate. At the present day, the Global South is generally thought to be a more suitable term for describing non-western or non-developed countries. It should be noted that the term Global South does not necessarily refer geographically to the south part of the globe: most people living in the Global South actually live in the Northern hemisphere (Hollington, Salverda, Schwartz & Teppe, 2015).

[2] Principles of Digital Development are widely used international principles for designing digital technologies to the contexts of developing countries. They were created by a broad group of NGOs and IGOs. More information at https://digitalprinciples.org/about/

## Theoretical framework: feminist technoscience and value-sensitive design approach

Our theoretical base here is twofold. It combines perspectives from feminist technology studies and value-sensitive design approach.  The former is used to analyze the results of case studies whereas the latter is mostly used to make policy recommendations. We will start by explicating the core ideas of feminist technology studies and feminist technoscience. We then move on to define our value-sensitive design approach. In the third section of this research paper, we will apply this general theoretical framework to mobility-related female emancipation.

### Feminist technoscience

Feminist technoscience is a critical interdisciplinary academic field of research (Åsberg & Lykke, 2010) that is interested in the ways gender and technology are mutually shaped (Wajcman, 2010). It is an academic stand supported by such researches as Judy Wajcman and Donna Haraway. Feminist technoscience (from here on FTS) is a feminist approach to science and technology studies (STS). The term technoscience is meant to be used to criticize the positivist distinction between scientific theories and their applications. Amongst researches of FTS and STS, it is seen that "pure and basic" science is tangled with societal interests (Åsberg & Lykke, 2010, p. 299).

Current feminist technology studies focus on the mutual shaping of technology and gender and conceptualize technology to be both a source and consequence of gender relations. These theories try to avoid determinism and gender essentialism and emphasize that "gender-technology relationship is fluid and situated" (Wajcman, 2010, p. 143). They highlight "how processes of technical change can influence gender power relations". The solution for gender equality in technology lies in a feminist view of technology policy (Wajcman, 2010, p. 143). We subscribe here to contemporary notions of feminist technoscience, but a few important points should be noted about the development of feminist theories of technology.

Initial feminist perspectives on technology were raised to criticize the masculinity of technology and the formation of engineering as a white and male middle-class profession which ruled out women and other groups. It is seen that "the hegemonic form of masculinity is still strongly associated with technical prowess and power" in Western society (Wajcman, 2010, pp. 143-145). This hegemonic position is built for instance through different childhood exposures to technology, the existence of different gender roles and by the segregation of job markets (Wajcman, 2010). Wajcman states that the marginalization of women from the technological community has had and keeps having a thorough influence on design, technical content and use of artefacts (Wajcman, 2010).

Subsequently in FTS the social factors shaping different technologies came under scrutiny. This especially happened from the point of view of how technology reflects gender divisions and inequalities in general, and how gender is embedded in technology itself (Wajcman, 2010). One can uncover different strands from this era of feminist STS, such as liberal feminism, radical feminism and socialist feminism. These stands have been criticized to be too pessimistic and too dismissal about women's agency as they emphasized "the proclivity of technological developments to entrench gender hierarchies" instead of their possibilities to yield change (Wajcman, 2010, pp. 146–147).

Contemporary approaches are more optimistic about the possibilities of ICTs to empower women and modify gender relations. Some cyberfeminists for example point out that digital technologies can blur the lines between humans and machines as well as male and female, making it possible to choose new identities and elect their disguises. These new digital technologies are based on different assumptions than industrial technologies. The internet and cyberspace are seen to be feminine media, possibly providing a basis for a new form of society which could be liberating for women. Things such as reproductive technology are fundamentally confronting more "traditional notions of gender reality" (Wajcman, 2010, pp. 147–148).

According to Wajcman, developments in digital technology call for a rethinking "of the processes of technological innovation and their impact on the culture and practices everyday life" amongst contemporary approaches (Wajcman, 2009, p. 148). For example, Donna Haraway elaborates a new feminist 'imaginary' which differs from the 'material reality' of the status quo technological order. Wajcman states that "to move forward, we need to understand that technology as such is neither inherently patriarchal nor unambiguously liberating" (Wajcman, 2010, p. 148).

FTS shares the idea that "technological innovation is itself shaped by the social circumstances within which it takes place" (Wajcman, 2010, pp. 148-149)[3]. Technology is treated as a sociotechnical product and as a "seamless web or network combining artefacts, people, organizations, cultural meanings and knowledge" (Wajcman, 2010, p. 149). A social constructivist framework is widely adopted amongst feminist STS scholars (Åsberg & Lykke, 2010; Wajcman, 2010) and following from this, "the gendering of technologies can then be understood as not only shaped in design, but also shaped or reconfigured at the multiple points of consumption and use" (Wajcman, 2010, p. 149). It follows that "gendered conceptions of users are fluid, and that the same artefact is subject to a variety of interpretations and meanings" (Wajcman, 2010, p. 150).

## Value-sensitive design approach

Value-sensitive design approach is an engineering methodology to integrate ethics, ethical responsibility and human values into the design of technology (Cummings, 2006; Friedman, Kahn & Borning, 2002). Value-sensitive design (VSD) is essentially

[3] Likewise, the concept of gender itself is also understood as "a performance or social achievement, constructed in interaction"; gender identities are shaped with the technology in the making, meaning that both gender and technology are product of moving relational process, emerging from collective and individual acts of interpretation" (Wajcman, 2010, 150).

a formalized scheme of technological design and engineering. It is a structured approach to incorporate human values and ethical concerns into the design process. Its design phases are similar to the typical system engineering approaches (Cummings, 2006).

VSD draws on moral epistemology and accounts for human values in the design process though a repetitive three-part design approach which takes conceptual, empirical and technical issues into consideration. VSD focuses on "broad, widely-held human values such as well-being, welfare and human rights", as opposed to the more personal values of individuals (Cummings, 2006, p. 702; Friedman et al., 2002, p. 1).

VSD emphasizes how technology shapes society and is being shaped by social factors. For this, technology cannot be made in a value vacuum as sociotechnical systems have intertwined with human-technology-interactions. There are twelve specific human values of ethical importance taken in consideration in the design process, such as human welfare, privacy, and environmental sustainability (Cummings, 2006.). These selected values are not independent or exclusive.

The conceptual investigation centers on the question of "how the relevant human values are either supported or diminished by a particular design" (Cummings, 2006, p. 702). It does not only contemplate those human values which could be supported or diminished by the particular technology. It also considers how "the technology could both socially benefit and negatively impact stakeholders"; potentially affected stakeholders should be considered both directly and indirectly (Cummings, 2006, p. 703; Friedman et al., 2002, p. 3). Often, especially indirect stakeholders — those affected by the system but not directly using them — are ignored in the design process (Friedman et al., 2002, 3).

The second phase is an empirical investigation that focuses on qualitative and quantitative measurements. The goal is to evaluate the design from a technical and value assessment approach. The most important consideration is what kind of effects design trade-offs have on "perceptions, behaviours and prioritization of competing values" and how the designer can contribute to or diminish value conflict (Cummings, 2006, p. 703). According to Friedman et al., "empirical investigations encompass any human activity that can be observed, measured, or documented" so the whole selection of "quantitative and qualitative methods used in social science research may be applicable here" (Friedman et al., 2002, p. 3).

The third step is to investigate technical issues. Different technical designs are analysed to determine how they can support particular values and in which ways the values identified in the conceptual phase can be supported in the most desirable way by different design possibilities. The difference between empirical and technical investigations is that the empirical investigation focuses on human-technology interaction whereas technical investigation is concerned with the technology itself (Cummings, 2006; Friedman et al., 2002).

### Conclusion

It should be noted that our way of applying value-sensitive design methodology might differ from the usual approach. This is especially the case in the sense that we are going to use it to analyse already-in-use-technologies to make general policy recommendations on the design process of ride-hailing apps. We argue that this is a useful approach when analysing the challenges related to mobility technologies and their consequences on human values such as female emancipation.

A second consideration is that we do not have direct knowledge or data about the design processes of these apps, and that is not really even our interest here. We are more interested in how these apps have been successful or unsuccessful in incorporating the values of VSD from an outside perspective. Has Uber been able to fulfil the values of trust, human welfare and privacy? Has Lyft worked for or against human rights? Are these ride hailing apps designed in ways that emancipate women?

## Contemporary development of mobility technologies and their potential emancipatory effect

Mobility, as in the ability to move freely, is an essential factor of emancipation. What is meant by emancipation here is the process of being set free from legal, social, or political restrictions. Mobility provides access to essential activities and enables people to "appropriate" their right to the city (Levy, 2013, p. 47). A moderate amount has been written about mobility from a gender perspective (see for example Riverson, Kunieda, Roberts, Lewi & Walker, 2005; World Bank, 2010). However, with the emergence of new technologies that may enhance mobility, a new area of research remains to be uncovered. In this section we aim to map out potential ways in which new mobility technologies could have emancipatory effects.

### Mobility and gender

Opportunities for mobility within and between cities relies strongly on transport systems. As transport systems are often planned primarily according to the travel needs of men (World Bank, 2010), women tend to have higher mobility constraints. Mobility and gender have a bilateral relationship: while mobility constraints clearly affect gender equality and the life opportunities of women, existing gender inequalities also have an effect on how transport is planned. According to the European Institute for Gender Equality (EIGE) (2016), transport and gender inequality intertwine and show in the following ways:

- access gaps to transport infrastructure and services
- segregation in transport labour market
- gender-based violence in transport
- weak representation of women in the decision-making processes in the transport sector

Women's mobility constraints are multiplied by the effect of social and cultural norms and practices. According to UNESCO, restrictions on women's mobility and access to public spaces can include limitations on the purpose and timing of their travel and controls of place, companions and their way of dressing (UNESCO, 2011). Women also tend to have a multifaceted role as workers, household managers and community managers, which often leads to so called time poverty (Venter, Vokolkova & Jaroslav, 2006). This gendered pattern of time-use is a clear hindrance to female emancipation, and it is enforced by the mobility constraints mentioned above. Studies have shown that gendered time poverty could be reduced with improved transport infrastructure (Asian Development Bank, 2015).

## Mobility enhancement through technologies?

Modern information technologies have enabled the emergence of a new kind of on-demand car services. Most of these new services function in the form of mobile applications that will from now on be called ride-hailing apps or services. Examples of such applications include Uber, Lyft and Ola. These services typically have improved productivity compared to traditional taxi services, mostly due to a more efficient technology to match the customer and the driver, and technologies that allow dynamic pricing (Rodrigue, 2017). The growth of ride-hailing services has also changed the ownership system of vehicles towards a leasing system, and increased the supply of driving services to meet the demand (ibid).

While ride-hailing apps have had a significant impact on the transportation industry for both its workers and customers, they also have implications for other sectors such as car manufacturers, insurance companies and telecommunication companies (Eisenmauer, 2018). The latter introduces an interesting thread of thought, since the quick emergence and growth of ride-hailing services has produced an unprecedented amount of passenger (and driver) data. So far, the data has mostly been used by the companies themselves to predict customer behaviour and improve customer experience. However, data collected by private companies can also be used to create and share knowledge on social issues such as gender gaps. One example of this kind of use of private company information is a report published by the International Financial Corporation (IFC) in cooperation with Uber and Accenture (IFC, 2018). We will use this report later on in this paper as material for our case studies in sections 4 and 5 and for our policy recommendations in section 6.

## Incorporating feminist technoscience and value-sensitive design to mobility technologies

As seen in feminist technology perspectives, technology can be seen as both a source and a consequence of gender relations (Wajcman, 2010). Incorporating this view of the issue of mobility constraints caused by transport (EIGE, 2016), an analysis can be formed. In the case of ride-hailing services, technology can be seen as a source of gender relations through either filling or enlarging women's access gaps, improving or diminishing segregation in the transport labour market, and possibly reducing or enforcing gender-based violence or harassment. The question of the representation of women in the decision-making of the transport sector in this context is a more complex one, as ride-hailing services have so far been ruled by private companies.

Whether gender-sensitive decision-making processes within the ride-hailing app companies could affect gender equality in a larger societal context is unfortunately out of the scope of this paper.

On the flip side of the coin, ride-hailing apps can be seen as a consequence of gender relations in the sense that the more women are empowered and encouraged to leave the domestic sphere, the more ride-hailing companies will have potential female customers and drivers. This could increase the amount of total supply and demand of ride-hailing, which would lead to increasing profits for the industry. From a more theoretical (feminist STS) perspective, the forms and dynamics of ride-hailing apps can be seen as a reflection of the social and cultural environment of the ride-hailing companies. Any gender imbalances or gaps can thus be associated with corresponding disorders in the surrounding world.

As we argued in the previous section, we believe that value-sensitive design may be a useful approach to solving potential gender issues in the ride-hailing industry. For example, if it is discovered that ride-hailing apps are rather enlarging access gaps than filling them in some contexts, VSD may be a beneficial way to approach the problem.

### Conclusion

To this day women face considerable mobility constraints that restrict the social, economic and cultural opportunity in their lives. Women's multifaceted role as household workers, community managers and sometimes workers, as well, causes a time poverty that is enforced by the mobility constraints. Emancipation requires better access to services, opportunities in the labour market and a safe mode to get around. While mobility constraints and the transportation industry limit the life of women, gender inequalities also affect the industry. This is to say that the relationship between gender inequalities and mobility and technologies is a fluid and bilateral one.

In our view there is considerable potential in the ride-hailing service industry to accelerate the process of female emancipation. To realize this potential, companies in the industry would have to consider the social aspects and impacts of their business in addition to the economic and environmental ones. One way of doing this would be through the application of value-design theory principles. Through effective and value-based decision-making processes and implementation in these companies, ride-hailing apps could have a significant effect even on policy areas such as health or economic development.

## Case study A: ride-hailing apps: driver's perspective

For the first case study of this research paper, we look at ride-hailing apps from the perspective of female drivers. In general, drivers who are women are quite rare when it comes to ride-hailing apps such as Uber or Lyft. Hence, data on female drivers is harder to find: for example, the IFC report used in this research paper deals with only 7357 female drivers (IFC, 2018). One of the most popular ride-hailing apps, Uber, reports that approximately 59,1 % of its drivers are male (Statista, 2019). The low number of female drivers is highlighted especially in the Global South — in Egypt, only 0,2 % and in Mexico 5,2 % of the Uber drivers are female (IFC, 2018). It also has to be noted that a lot of research conducted in general about drivers using ride-hailing apps is solely about male drivers (e.g. Kashyap & Bhatia, 2018). The resources on female drivers are hence rare, and because of this, the following section of our research paper deals mostly with female drivers in general, without any strict geographic demarcation.

Female drivers in ride-hailing services work as drivers for a number of reasons. In countries of the Global South such as Egypt, the main reason is to boost their income (IFC, 2018). Ride-hailing apps make it possible for women to choose their own working days and hours, which makes it easier for them to work as drivers in addition to performing household duties (ibid). These findings go along with our theoretical framework: ride-hailing apps represent new digital technologies that make it possible for women to find new areas of work and lead to emancipating effects. Apps such as Uber suit the living conditions of women who are thus not solely dominated by patriarchal conditions.

### Driving and emancipation

As noted in Section 2, contemporary feminist theories are positive in regard to the emancipatory effects of ICTs. New "digital technologies can blur the lines between humans and machines, and male and female" (Section 2). For example, Uber has been described as a "gender-blind" tool as its payment method is solely based on the driver's work conducted and not gender (Cook, Diamond, Hall, List & Oyer, 2019). In general, new ICT technologies can be seen to emancipate women (Huyer & Sikoska, 2003). The biggest benefit that women can receive from new ICT technologies comes from the information that women can have access to (ibid). When it comes to female drivers, the emancipatory effect is, however, limited.

As observed above, for female drivers these ride-hailing apps offer the potential of earning more, as well as an opportunity to work around other duties (IFC, 2018). The average income rise is 13% for women, whereas for men it is only half of this (7%). Many women who start working as drivers do not have a full-time job (or any job) before signing up, which also explains the big surge in their income. The importance of working as a driver for women's income cannot be understated, and it can lead to emancipation, as women have more control over their own income. For many women, working as a ride-hailing driver might be the first opportunity to make money on their own. (Ibid)

Female drivers on the ride-hailing apps can be described as "pioneers", as they are challenging the dominant cultural norms by choosing to work as drivers (IFC, 2018). IFC notes that "51% of women drivers decided to sign up with Uber because a friend or a family member suggested it" (ibid, p. 35). As such, working as a driver can have a positive effect on the surroundings of the "pioneer", and encourage other women towards empowerment by setting an example.

It is also important to note that ride-hailing apps offer easily accessible job opportunities in countries where there is a large body of young and unemployed workforce. In this way, it could also be argued that ride-hailing apps emancipate women by helping to create conditions for a more secure society by offering low-threshold job opportunities.

## Obstacles for the female driver's emancipation

However, there still is a considerable set of obstacles for women's emancipation as ride-hailing app drivers. As noted in Section 2, ICTs are not invented in a vacuum; they reflect the society which they were created in. The unequal and gendered structures present in social relations also map onto technology and as such technological developments reflect the world in which they were created.

For female drivers, safety issues represent a big obstacle for working freely. In both the Global South and the Global North, women face threats of harassment and violence when they work as drivers (Brenton & Curry, 2016; *The Guardian* 19th of June 2019; IFC, 2018). In Saudi Arabia, where driving was forbidden for women until 2017, female drivers and activists on the subject have been persecuted (Amnesty International, 2018). The social norms such as "women should not be driving" are still strong in some parts of the world, and women who work as drivers have to overcome social stigma (IFC, 2018). For example, in Egypt "57% of men would be unhappy if a female family member wanted to sign up" for Uber (ibid, p. 34).

In terms of emancipation, it also has to be noted that most women who drive are doing so because they are the main provider of the household's income. In other words, their driving is due to financial reasons and does not necessarily cause emancipation (IFC, 2018). Working as a driver might lead to emancipation, but it might also stem from a need to work rather than the desire for it. In addition, when it comes to ride-hailing apps being a tool for women to build their own businesses, this was only partly supported by the data offered by the IFC 2018.

Ride-hailing apps so far have failed to bridge the gender earning gaps. Drivers are paid according their experience and the routes they take, and as such ride-hailing apps could be assumed to be equal and even "gender-blind" tools. However, a study has shown that there is in fact a 7% gap between the earnings of male and female drivers (Cook et al. 2019). This comes mainly from three factors: "returns to experience, a pay premium for faster driving and preferences for where to drive" (ibid, p. 38).

Nevertheless, all of these three factors point the blame more or less towards the pre-existing status of women. Women have fewer opportunities to choose when and where they drive because of the safety issues mentioned earlier. According to the IFC report, it is also common that passengers might avoid female drivers because of their gender (IFC, 2018). As a result, gaining experience on the platform is harder for female drivers than it is for men. It is also false to assume that Uber would be a gender-blind application: passengers are able to see the name and photo of their driver when they are ordering their ride (ibid).

### Conclusion

In conclusion, the emancipatory effect for female drivers is quite limited. This cannot be stated with a lot of confidence yet, as only little research has been conducted on the matter. More research, especially on female drivers in the Global South should be conducted before making solid conclusions. Next, we turn to female passengers' perspective in using ride-hailing apps in the global South.

## Case study B: ride-hailing apps: passenger's perspective

According to the World Bank, gender-based inequalities in transport in developing countries result in slowing economic growth and poverty reduction (World Bank, 2010). Transportation provides access to many empowering features for women: employment, education, health, childcare and political participation (ibid). Thus, the focus in this case study is on the liberating effects ride-hailing apps can offer for their women passengers. First, we briefly introduce the basic structures of mobility in developing countries and who constitute the women passengers' group, then we discuss the three potentially emancipatory effects of ride-hailing apps: safety, mobility, and affordability, and we finish with conclusions.

### Mobility in the developing countries

In many developing countries rapid urbanization has increased the demand for mobility services, which public officials have been unable to meet with formal public transportation services (Kumar, Seema, Akshima, Sarbojit & Wilson, 2016). Thus, the informal transportation provided by non-government actors is crucial for the population to meet their transportation needs (ibid). Informal transportation usually consists of shared buses, taxis, or bicycles, but privately-owned cars operating via the new ride-hailing apps can also be included in this category, even though there are significant differences in the pricing of these services.

The IFC report suggests that women face more constraints on their ability to travel compared to men, and this limits their development and social mobility (IFC, 2018). New technological innovations such as ride-hailing apps have offered new opportunities for women to participate in the economy, and increase their social and economic autonomy (ibid).

### Women passengers

In this case study, we use data and examples from three countries to evaluate the possible emancipatory effects of ride-hailing services. These countries are India, Mexico, and South Africa. The main source of our data is the International Financial Corporation's report based on Uber's data on women and ride-hailing.

In India 31%, in Mexico 47%, and in South Africa 45% of Uber riders are women (IFC, 2018). According to the IFC study, in Mexico 82% of female passengers are unmarried, and in India 72% of female passengers have an income above the median. The average ages these riders vary from the 26 years in Mexico to

36 years in South Africa. (ibid) According to this data, Uber riders are hence relatively young and financially stable.

When discussing the emancipatory effects in the Global South, it is important to acknowledge that the proportion of women using ride-hailing services represents only a very limited group of the whole population. In developing countries walking is the predominant way of getting around (World Bank, 2010). Also, ride-hailing services are available in very few urban areas. Thus, the question remains: are the women that can be interpreted to be somewhat emancipated by the ride-hailing apps truly the group that needs emancipation? This is a normative question beyond the scope of this paper. Next, we will discuss ride-hailing apps' possible emancipatory effects.

**Emancipation and its implications**

The emancipatory effects of ride-hailing apps in the Global South can be divided into three categories: safety, mobility, and affordability. Ride-hailing apps have increased the traceability and the predictability of the means of mobility, and these features can be argued to increase the safety of the users. In Mexico 49% and in South Africa 38% of female riders identify the security features of the Uber app as being particularly important. 51% of the South African women riders identify knowing the driver's name in advance as a key benefit (IFC, 2018).

Increased safety may result in women's emancipation, but the security features can also affect societies by instrumentally increasing trust among people. The development of the sharing economy increases trust in society by facilitating new ways for unfamiliar people to cooperate. This happens via peer-to-peer review systems, which enable the documentation of who is trustworthy (IFC, 2018; Schoenbaum, 2016).

On the other hand, the intimate nature of sharing economy transactions also creates multiple safety threats for its users. For example, many cases of Uber drivers sexually harassing their passengers have been reported and made public (Schoenbaum, 2016). As a result, many female drivers and riders have expressed preferences to transact with other females (ibid). According to predictions, an increase in the number of women drivers would probably also increase the number of women riders (IFC, 2018). There are a few women-only ride-hailing apps in India, Mexico and the US (IFC, 2018), but these apps have been under discussion for being discriminatory or even illegal (Tarife, 2017).

Uber has, however, tried to acknowledge some of the security issues. For example, in South Africa, they established a call-back line for safety-related matters (IOL, 2018). Uber also increased their private security response teams in key risk areas, such as Gautrain stations in Johannesburg area (ibid).

As stated in the previous sections, women have complex mobility needs due to their roles as household managers, community managers, and workers (Venter, Vokolkova & Jaroslav, 2006). The second emancipatory effect relates to improving women's mobility in general by offering more flexible mobility solutions. The IFC study reveals that women are using ride-hailing apps to take care of household or social issues (IFC, 2018). Compared to public transportation, female riders emphasize how Uber facilitates the speed and independence of their mobility. A significantly larger proportion of women compared to men also use the ride-hailing when traveling with children (ibid).

Women's multiple roles and their complex mobility needs often lead to time poverty, as stated in the third section (Venter et al., 2006). One way to reduce this is the improvement of transport infrastructure (Asian Development Bank, 2015), to which ride-hailing apps can be considered to contribute.

The third emancipatory effect of ride-hailing is cost transparency and affordability. According to the IFC study, knowing the cost of a ride in advance is important for the women riders (IFC, 2018). Ride-hailing has also resulted in lowering the transportation prices of private taxis in many regions caused by the demand-based pricing Uber and other apps use (O'Toole & Matherne, 2017).

Yet, it is necessary to acknowledge the exclusivity of ride-hailing apps compared to the informal public transportation means used by the majority of the population. This notion was also acknowledged in the IFC report, where it was stated that the price of ride-hailing services still limits their use. The report recommends developing lower-cost models and supporting more affordable options such as motorcycles, so that less affluent segments of the population can also access these ride-hailing services. (IFC, 2018) This notion will be discussed further in the following section of this paper.

### Conclusion

As noted above, ride-hailing apps in the Global South offer many promises for women's emancipation, such as an increase in the safety of the passengers, more flexible mobility services, and affordable and cost-transparent pricing. However, all of these effects require more research before further conclusions can be stated. In the Global South walking and affordable informal public transportation play a major role in fulfilling people's mobility needs, and thus the group of women potentially emancipated by ride-hailing apps remain relatively limited.

## Policy recommendations based on the principles for digital development

In this section we will provide some policy recommendations for the design and further refinement of ride-hailing apps, so that they would take into consideration gender-technology-relations and their effects on women's mobility in the Global South. Recommendations are based on the Principles for Digital Development and our overarching framework of Value-Sensitive Design (VSD).

### The principles for digital development

Following the VSD methodology, the first step in the evaluation of the ethical consequences of technological design is a conceptual investigation. As elaborated above, VSD mandates the consideration of twelve human values in the making of a technological design, including human welfare, ownership and property, privacy and universal usability. The conceptual investigation examines what the most critical human values to take into consideration are in the designing process, and whether they are supported or undermined by it (Cummings, 2006).

We argue that especially the values of *human welfare* and *universal usability*, as formulated by VSD, are central ethical issues when designing and developing ride-hailing apps. These somewhat broad values can be refined by applying the Principles of Digital Development (hereafter Principles), which are guidelines for applying digital technologies to development programs, formulated in cooperation by several international development organizations, such as The Bill and Melinda Gates Foundation, UNICEF and the World Bank (Principles of Digital Development, 2019). Currently, the Principles consist of nine guidelines: design with the user; understand the existing ecosystem; design for scale; build for sustainability; be data driven; use open standards, open data, open source, and open innovation; reuse and improve; address privacy and security; and be collaborative (ibid).

Although ride-hailing apps certainly cannot be classified as development programs, we nevertheless feel that the Principles can be used as a guide in evaluating their gendered effects in the Global South. Based on our case studies and the feminist theoretical framework, we argue that the principles of *designing with the user*, *understanding the ecosystem* and *reusing and improving* are the most relevant guidelines to look into when adopting digital innovations in developing regions. However, it must be noted that the Principles are not clear-cut and instead overlap in many ways, and that a genuinely successful developmentally oriented initiative must consider them all. Below we will elucidate what the selected principles encompass.

The principle of user-centered design recommends getting to know future users through conversation, observation and co-creation as a starting point for designing technological programmes. The information garnered in this process is to be used in the building and testing of tools until they meet the need of the users. This also includes continuously consulting different user types and stakeholders and incorporating their feedback in further developing the programmes. It instructs to develop context-appropriate tools and ensure that the design considers the needs of the traditionally underserved, such as rural women in developing countries (Principles of Digital Development, 2019).

Understanding the ecosystem entails a continuous dedication of time and resources to get to know the operational environment, and make sure that the technological innovations are relevant, sustainable and sensitive to the existing ecosystem. Ecosystems comprise of a large scale of factors, such as the culture, economy and political landscape, that can constrain individuals' access and usage of technological tools. The user-centered principle's advice is not only to engage in a dialogue with users, but also it mandates a coordination with larger organizations to avoid duplicating efforts and to integrate with the existing systems. When applying this principle, one should involve community members, governments and other affected bodies throughout the program's lifecycle. Getting to know the ecosystem and the users involves gathering data about the context and the stakeholders who are directly or indirectly affected by the program (Principles of Digital Development, 2019).

The principle of reusing and improving refers to evaluating available technologies and resources and building on top of them to make them meet the needs of the ecosystem, rather than creating something new from scratch. It begins by identifying the existing technological tools, by for example collaborating with other actors in the field, and then evaluating if they can be reused or modified in the current scenario. (Principles of Digital Development, 2019.)

### Policy recommendations for the design of ride-hailing services

Next, we will share some examples of policy recommendations that draw on the principles introduced above, and which seek to address the problems that women face in the ride-hailing industry, identified in our case studies. When exporting services of Western origin to the Global South, companies should familiarize themselves with the users of the service and the ecosystem, as the operational environment of the area, and utilize or build on already existing techniques where appropriate. Some of the policy recommendations draw on the IFC report (2018) that was utilized earlier in the case studies, but from the perspective of the selected principles of digital development.

Working for ride-hailing companies can enhance women's financial independence and therefore lead to increased emancipation. As mentioned in section four, female drivers are often working as pioneers and encourage other women to pursue independence through working for ride-hailing companies. This could be further enhanced for example by opening "pop-up" recruitment booths in areas accessible to women, by simplifying driver registration processes, or by creating online peer support groups for female drivers (IFC, 2018).

For passengers a key factor in the deployment of ride-hailing services is cost transparency and affordability. The industry could open ride-hailing to lower-income populations and thereby enhance their access to mobility through expanding ridesharing options or providing options that would trade-off the lower price to other aspects, such as higher travel times (IFC, 2018).

As our case studies pointed out, possibly the most critical factors for women in using ride-hailing services are safety related issues. Women face more risks of gender-based violence in transport. They adapt their driving and riding patterns to alleviate those risks, and by doing so, they reduce their mobility and income.

Consistent with the Principles, we suggest that in order to alleviate these safety risks, companies should draw on the most advanced security features from across industries, for example by deploying the latest techniques in ways to identify the driver (such as visual recognition). Companies should also partner up with other larger organizations, such as law-enforcement agencies, to facilitate the coordination in situations of emergency. Ride-hailing companies could engage with the female users actively and habitually to learn about the nature of security, what kind of threats are perceived, and perhaps through online forums. Feedback from the users could be incorporated into the improvement of the services.

Another manifestation of unequal gender status in ride hailing apps is the earning gap between male and female drivers, as articulated in the first case study. Uber's practice of showcasing drivers' photos and names can lead to women gaining less experience on the platform and through that to an augmentation of gender earnings (Cook 2019). Trade-offs like this need to be systematically reflected on in the second phase of VSD, that contains an analysis of how different trade-offs can result in value-conflict. Ride-hailing companies can affect the earning gap also by working with policymakers and other sharing economy platforms to implement a system of portable benefits drivers could carry across platforms and apps, or by collaborating with financial institutions to develop insurance and pension products tailored for independent contractors (IFC, 2018).

### Conclusion

Ride-hailing services have the potential to have an emancipatory effect for women (and other marginalized groups too) in the Global South. However, companies must acknowledge that a design of Western origin may not work in other regions as such. We have argued that for ride-hailing services to be truly successful and ethically sustainable in the Global South, companies must take into consideration the Principles for Digital Development and especially the guidelines that mandate the consideration of users and the ecosystem, and suggest reusing and improving existing techniques. VSD could serve as a formalized methodology for companies in the integration of ethics into their services.

As stated above, recruiting more female drivers into ride-hailing apps can lead to a virtuous cycle by attracting even more women riders. Thereby finding ways to hire more women into the industry, the emancipatory effects could cumulate. We think that following the policy recommendations suggested above when crafting new policies and practices in the industry could therefore be profitable to ride-hailing companies, as this would increase their number of customers. However, more data on the ways women use ride-hailing services is necessary before establishing more sustainable practises in the industry. Therefore, gathering more data is our first and foremost policy recommendation.

## Conclusion

In this research paper we intended to study the emancipatory benefits ride-hailing apps such as Uber and Lyft can have for women in the Global South. We drew on the theoretical framework of Feminist Technoscience to analyze the interconnectedness of mobility and emancipation and used and a framework of Value-Sensitive Design to analyze ride-hailing apps. We took into consideration the perspectives of both drivers and passengers for us to be able to better evaluate the emancipatory effect in practice. The final section of our paper consisted of the Policy Recommendations based on the

Principles of Digital Development. Coming up with these recommendations, we found Value-Sensitive Design theory especially pertinent.

The paper was quite limited in its length and the case studies were able to offer only introductions to the wide themes under scrutiny. Thus, the results are not unambiguous. The conclusion of the first case study, the drivers' perspective, was that the emancipatory effect was quite limited. This was caused partly by the fact that there was only a limited amount of data available on women drivers. However, it was suggested that female drivers are emancipated by being able to access more income via a flexible work situation and that this might have an inspiring effect on their surroundings.

In the second case study concerning the female passengers, more emancipatory effects were found, although the data was similarly limited, and thus the results are not by any means fully applicable. The emancipatory effects of ride-hailing apps on women in the developing countries were divided in three categories: safety, mobility and affordability.

In the policy recommendation section, we shed light on the Principles of Digital Development and assessed how these can be utilized as a guideline in applying the VSD methodology in the pursuit of integrating ethics into design projects. We argued that ride-hailing companies should research the users and the ecosystem and utilize already existing suitable techniques when exporting services of western design and use VSD methods in product development. We provided a few examples of policy recommendations consistent with the principles.

We found multiple fascinating ways in which ride-hailing apps might have emancipatory effects. For future research we suggest a deeper investigation of how different context factors can be taken into account in the design processes of ride-hailing services.

## References

Asian Development Bank (2015). *Balancing the Burden? Desk review of women's time poverty and infrastructure in Asia and the Pacific.* Mandaluyong City, Philippines: Asian Development Bank.

Amnesty International (2018). *The driving ban and women's rights in Saudi Arabia*. Retrieved October 21, 2019, from https://www.amnesty.org/en/latest/news/2018/05/the-driving-ban-and-rights-in-saudi-arabia/

Cook, C., Diamond, R., Hall, J., List, J. A., & Oyer, P. (2019). The Gender Earnings Gap in the Gig Economy: Evidence from over a Million Rideshare Drivers. *NBER Working Paper Series,* p.24732.

Cummings, M. (2006). Integrating Ethics in Design through the Value-Sensitive Design Approach. *Science and Engineering Ethics 12(4),* 701–715.

Eisenmeier, S. R. (2018). *Ride-sharing Platforms in Developing Countries: Effects and implications in Mexico City.* Retrieved from https://pathwayscommission.bsg.ox.ac.uk/sites/default/files/2019-09/ride-sharing_platforms_in_developing_countries.pdf

European Institute for Gender Equality (2016). *Gender in transport*. Luxembourg, Luxembourg: Publications Office of the European Union.

Friedman, B., Kahn, P. & Borning, A. (2002). Value Sensitive Design: Theory and Methods. University of Washington: *UW CSE Technical Report 02-12-01*, 1-8.

The Guardian. 19th of June 2019. *Female drivers feel abandoned by Uber and Lyft after reporting a sexual assault.* Retrieved October 18, 2019, from https://www.theguardian.com/technology/2019/jun/19/uber-lyft-female-drivers-sexual-assault

Hollington, A., Salverda, T., Schwarz, T. & Tappe, O. (2015). *Concepts of the Global South.* Cologne, Germany: University of Cologne, Global South Studies Center Cologne.

Cook, C., Diamond, R., Hall, J., List, J. A., & Oyer, P. (2019). The Gender Earnings Gap in the Gig Economy: Evidence from over a Million Rideshare Drivers. *NBER Working Paper Series,* p.24732.

Cummings, M. (2006). Integrating Ethics in Design through the Value-Sensitive Design Approach. *Science and Engineering Ethics 12,* 701-715.

Huyer, S., & Sikoska, T. (2003). Overcoming the Gender Digital Divide: Understanding ICTs and their Potential for the Empowerment of Women. *Instraw Research Paper Series 1.*

IOL (2018). *Uber introduces rider call back line for added safety*. Retrieved October 18, 2019, from https://www.iol.co.za/travel/travel-news/uber-introduces-rider-call-back-line-for-added-safety-13756765

International Financial Corporation (2018). *Driving toward equality: Women, ride-hailing and the sharing economy.* Retrieved from https://www.ifc.org/wps/wcm/connect/62a2871b-271b-4256-b426-

65b2012d00f7/00418+IFC+DTE+Report_Complete_Layout+Final2-pxp.pdf?MOD=AJPERES&CVID=m9ksr4q

Kashyap, R. & Bhatia, A. (2018). Taxi Drivers and Taxidars: A Case Study of Uber and Ola in Delhi. *Journal of Developing Societies 34(2)*, 169–194.

Kumar, M., Seema S., Akshima T.G., Sarbojit P. & Sangeetha A.W. (2016). Informal Public Transport Modes in India: A Case Study of Five City Regions. *IATSS Research* *39:2*, 102-109.

Levy, C. (2013). *Travel choice reframed: "deep distribution" and gender in urban transport*. Environment and Urbanization 25(1), 47–63.

Malin, B. J. & Chandler, C. (2016). Free to Work Anxiously: Splintering Precarity Among Drivers for Uber and Lyft. *Communication, Culture and Critique 10(2)*, 382–400.

O'Toole, J. & Matherne, B. (2017). Uber: Aggressive management for growth. *The CASE Journal,* 13:10.

Principles of Digital Development (2019). Retrieved October 26, 2019 from https://digitalprinciples.org/

Riverson, J., Kunieda, M., Roberts, P., Lewi, N. & Walker, W. M. (2006). Gender Dimensions of Transport in Developing Countries: Lessons from World Bank Projects. *Transportation Research Record Journal of the Transportation Research Board* *1956(1)*, 149–156.

Rodrigue, J. (2019). *Future Transportation Systems*. Retrieved from https://transportgeography.org/?page_id=1579

Schoenbaum, N. (2016). Gender and the Sharing Economy. *Fordham Urban Law Journal*; *GWU Law School Public Law Research Paper 53; GWU Legal Studies Research Paper 53.*

Statista (2019). *Distribution of Uber's employees worldwide from 2017 to 2019, by gender.* Retrieved October 21, 2019, from https://www.statista.com/statistics/693807/uber-employee-gender-global/

Tarife, P. M. (2017). Female-Only Platforms in the Ride-Sharing Economy: Discriminatory or Necessary. *Rutgers University Law Review, 70:295.*

UNESCO Office New Delhi & Centre de Sciences Humaines (India) (2011). *Urban policies and the right to the city in India: rights, responsibilities and citizenship.* New Delhi, India: UNESCO New Delhi/CSH.

Venter, C., Vokolkova, V. & Jaroslav, M. (2006). Gender, residential location, and household travel: Empirical findings from low-income urban settlements in Durban, South Africa. *Transport Reviews 27(6)*, 653–67.

Wajcman, J. (2010). Feminist Theories of Technology. *Cambridge Journal of Economies 34(1*), 143–152.

World Bank (2010). *Mainstreaming Gender in Road Transport: Operational Guidance for World Bank Staff.* Washington, DC: World Bank.

Åsberg, C. and Lykke, N. (2010). Feminist technoscience studies. *European Journal of Women's Studies 17:4,* 299–305.

# 4.2 Human Robots as Members of Future Society? — The Case of Sophia

Kaarlo Somerto, Elisa Seppänen, Helmi Rantala, Essi Pitkänen, Eeva Nyman
Faculty of Social Sciences, University of Helsinki

(ITU Pictures. Retrieved from Wikimedia Commons)

## Abstract

This paper discusses the role of human robots in society. We are looking at this phenomenon through Sophia, the human robot that has gained citizenship status and notable attention. In the first section we ask if Sophia, and human robots more generally, should be referred to as "it" or "she" by considering the human perception of a human robot's appearance. Second, we consider the rationality and power relations behind human robot action. Third, robot rights are examined in relation to human rights. Drawing from these discussions, we argue that introducing human robots into the workforce portrays the tendencies of a society to be based on instrumental relationships between humans. Finally, we consider the implications of delegating political decision-making from humans to human robots as a form of AI. As artificial intelligence is increasingly developing, it is important to consider its legal, ethical and societal implications.

*Keywords*: Artificial intelligence, AI, human robots, ethics, human rights, human-robot relations, robot citizenship, Saudi Arabia, citizenship, Japan

The digital revolution—the growing prevalence of technology and digitalisation in all spheres of life—has caused humans to rethink their role in the world (see summarization in Wessels 2007, p. 1-13). In this paper, we are looking at the implications of artificial intelligence and human robots on public life through the case of Sophia, the human robot. Sophia is an interesting example of this development because she has been granted human-like rights, for example citizenship in Saudi-Arabia. Sophia is the newest and the most advanced human robot of a Hong Kong-based technology company, Hanson Robotics Limited (UNDP 2017).

Sophia, as many human robots, utilizes artificial intelligence (AI), which Vincent Boulanin (2019, p. 14) defines as a "catch-all term that refers to a wide set of computational techniques that allow computers and robots to solve complex, seemingly abstract problems that had previously yielded only to human cognition". The questions of robots' societal role are increasingly significant with the advancement in machine learning and thus hypothetically the possibility of robot autonomy. Machine learning can be understood as "an approach to software development that first builds systems that can learn and then teaches them what to do using a variety of methods (i.e. supervised learning, reinforcement learning or unsupervised learning)" (Boulanin 2019, p. 15). Machine learning has the potential of making robots autonomous from their programmers and can therefore have unforeseeable impacts on technology, human-robot relations and society. The recent advancements in robotic technology bring science fiction closer to our everyday lives, as in the example of the human-like robot Sophia.

In this paper, we address questions such as whether human robots can be independent members of the public sphere and participate meaningfully in public discourse. If not, who is the actor holding power behind human robots? What are the implications of a human robot or AI participating in decision-making? We consider the human perception of human robots through their appearance. In addition, we address the kinds of rights human robots could obtain in relation to human citizens. These questions must be examined as we enter a new era in which artificial intelligence and human robots will inevitably continue to have an impact on our societies.

## It or She? Human Robot Appearance and Human-Robot Relations — Kaarlo Somerto

In this section I consider whether Sophia the human robot should be addressed using the pronoun "it" or "she". This consideration may seem fairly superficial at first glance. However, when contemplated further, it leads us to consider fundamental questions surrounding the agency of Sophia and human robots more generally.

In the very early, but still influential account of AI and machine learning, Alan Turing (1950) defined AI not through its intellectual qualities per se, but through how people perceive its intellectual qualities to be –i.e., how well an AI can "imitate" a human mind (Turing 1950, p. 433). The same logic is applied here to the account of Sophia's appearance: I am interested not in Sophia's technological properties and details, but in how well it achieves its stated goal of being human-like (Hanson Robotics 2019b) in that people see Sophia as human-like.

Human perception of a human robot is affected by verbal (Turing 1950) and nonverbal communication, along with other physical aspects of the robot's appearance, such as how it interacts with its surroundings through movement (Alač et al. 2011), and its looks, e.g. having a "human face" (Goertzel et al. 2017, p. 3). Drawing on these considerations, what is meant here by human robot appearance is: how human-like do humans see a human robot, based on the aforementioned aspects of its appearance.

The division between the use of "it" and "she" when referring to a human robot—i.e. the it-she divide—is very pertinent in the case of Sophia. It is categorically referred to as "she" in communications by its creator, Hanson Robotics (Hanson Robotics 2019b), by Saudi Arabia (Center for International Communication [Saudi Arabia] 2017), and by some media outlets (e.g. Greshko 2018). On the other hand, other actors, such as certain other media outlets (e.g. Edwards 2017), experts on AI (Sharkey 2018), and even Sophia's AI's main developer Dr. Ben Goertzel (Russon 2019) have referred to Sophia as "it".

The it-she divide takes us to the deep and fundamental question of what it means to be considered to be human, as "she" clearly has a more human connotation than "it", serving as a means of humanizing the object to which it refers. While Sophia indeed has been referred to as both "it" and "she", it is still, in November 2019, fair to say that the question of the actual division between human and robot is not ambiguous. The level of AI as well as robot technology more broadly has yet to reach a level where one could seriously contemplate whether something constructed to be a robot could be considered to be or become a human being, or whether someone born a human being could be or become a robot. Sophia exemplifies this: despite being described as "human-like" (Hanson Robotics 2019b), and despite having an unarguably human-like face as well as, to some extent, a human-like style of communication, it is certainly not indistinguishable from a human being. Sophia does, however, set an interesting precedent: with its human-like physical features and AI, developed further in the future, coupled with developments in robotic additions or alterations to the human body, humanity may face a future of a fluid continuum between human and robot, instead of the still presently existing stark divide. The mere fact that Sophia is in many instances called "she" instead of "it" shows that despite its "flaws", which make it distinguishable from a human, it has already reached a level of similarity to humans that raises questions about relations and divisions between humans and robots.

The question of how Sophia's appearance and interactions affect its perception by humans is entwined with the more specific question of how Sophia's social interactions affect human perception of Sophia and its intelligence. The ability of robots to interact with humans socially, in a socially intelligent manner, is and will continue to be a key dimension of human-robot interaction and relations (Dautenhahn 2007). No matter how intelligent or human-looking Sophia would be, it can be assumed that if it fails to communicate in a human-like fashion (if it for example continuously gives nonsensical answers to questions or moves in a rigid, not human-like, "robotic" way), it will not be accepted as *human-like enough* to "deserve" to be called "she", alienating itself in the eyes of humans from human-likeness. Sophia is often referred to as "she" by a variety of actors in the public sphere, which means that these actors (consciously or not) accept Sophia to be human-like enough to be deserving of being called "she".

According to researchers working on Sophia, Sophia is intended to be "loving" in its interaction (Goertzel et al. 2017, p. 2). This is meant to "help humans achieve greater states of well-being" (ibid.). This could be translated to mean that Sophia is intended to act as a robot version of a psychologist or compassionate friend—only infinitely patient. During their experiments wherein Sophia had social interactions with human beings, these researchers found that Sophia could indeed have a positive effect on people's emotions through social interaction (Goertzel et al. 2017, p. 7-8). People participating in the study reported that Sophia was "*human-like enough* to make your mind feel that another *person* is there, seeing you, mirroring you, paying close attention to you" (Goertzel et al. 2017, p. 10, emphasis added). In addition, the participants referred to Sophia as either "Sophia" or "she" (ibid.). Based on this, it is fair to say that Sophia's appearance can (at least in the circumstances of the laboratory) make human beings see it as human-like enough to not only call it "she", but to identify with it, and to experience positive emotions as a result of interacting with it.

Based on their study, Sophia's developers have expressed the view, among others, that "the AI/robot can express unconditional *acceptance* of people, in some ways/cases better than people can" (Goertzel et al. 2017, p.12, emphasis original). Based on their work with Sophia, these researchers believe that in the not so distant future robots with developed AI will be able to perform compassionate social interactions better than humans on a mass scale (Goertzel et al. 2017, p. 4). The researchers also claim that "AI can give people the experience of *being seen*… This has to do partly with physical interactions with the robot, like facial expression mirroring" (Goertzel et al. 2017, p. 12, emphasis original). This highlights the importance of the interplay between the robot's AI and its physical aspects: both must be sufficiently human-like. These notions also raise questions about the future of human-robot interaction and robot membership in future societies: should all or most people have infinitely patient robot companions to interact with, and how would this impact human-human relations on a mass scale and society as a whole?

Despite all of this, I personally remain committed to calling Sophia "it" instead of "she". Despite this section's argument being based on human perception of Sophia's appearance, rather than its actual qualities, I ground my decision in a moral and essentialist rationale. Despite great technological advancement, robots—even Sophia—are still ultimately dependent on their programmers. As long as artificial general intelligence (AGI) (an AI that would match or outperform "a human's ability to make sense of the world and to develop an understanding of its environment" (Boulanin 2019, p. 13)) remains unrealised, there remains a clear division between human and robot. Even in the case of AGI becoming reality, annulling that division would be dubious. The dichotomy of “it” and “she” exemplifies this division. I call Sophia “it” simply because I do not consider it human, nor a living being for that matter. Sophia is a machine, surely with human-like traits, but built and programmed by humans, and while able to evoke real feelings in human beings, itself non-sentient.

## Human Robots and Power in Public Discourse — Helmi Rantala

Sophia, the human robot, has made numerous public appearances—its whole essence seems to be publicity. In fact, it could be asked whether Sophia even has a private life: would that be the time spent on the programming, maintenance and repair of the robot? For instance, Sophia has its own social media accounts where the form of first-person singular is used. It has attended many popular television shows (e.g. The Tonight Show Starring Jimmy Fallon 2017), participated in skits (e.g. Smith 2018), held public speeches at influential events (e.g. Arab News 2017) and has performed in the UN (e.g. United Nations 2017). Sophia can also be booked to speak at an event, with prices ranging from $7,500 to more than a hundred thousand dollars (e.g. APB 2017).

Altogether, Sophia has gained an audience of at least tens of millions of people. Thus, it could be suggested that it is present in the public sphere. The question is whether a human-like robot, like Sophia, can act in the public sphere and public discourse as any human being. If Sophia is not an independent actor in the discourse, who benefits from its performances? Thus, if Sophia is not promoting its own agenda, whose agenda is being promoted?

In his text, “The Public Sphere: An Encyclopedia Article” (1964), Habermas defines “public sphere” as a “realm of our social life in which something approaching public opinion can be formed.” He continues by stating that access to the public sphere is guaranteed to all citizens. (Habermas 1964, p. 49) Technically, Sophia is a citizen of Saudi Arabia (e.g. Griffin 2017). However, is Sophia capable of participating in the public discourse in the same way as human citizens?

One way to approach the question of whether Sophia is an independent actor is to examine Habermas’ theory of rationality. Habermas sees the practice of rationality behind all action and language to be a key aspect of all contemporary philosophy (Habermas 1984, p. 2). For Habermas, there are two kinds of rationalities through which one can act: communicative rationality and instrumental rationality. While instrumental, strategic rationality seeks success for oneself, communicative rationality tries to build understanding (Schaefer i.a. 2013). Habermas for instance criticizes Weber for viewing all social rationalization through purposive rationality (Habermas 1984, p. 273-280). In communicative rationality, communicating is an intrinsic value itself, while in instrumental rationality communicating is seen as an extrinsic value (e.g. Zimmerman & Bradley 2019). While communicative rationality is not sufficiently encompassing to explain all societal actions, a society wherein understanding is put to the center would be ideal in many ways (Habermas 1989, p. 1-2).

Can an interaction with a human robot ever be communicative? Are human robots simply following the orders that they have gotten from a programmer? It is easier to understand that artificial intelligence and machines function through instrumental rationality if the machine's appearance does not resemble human like features like Sophia. The most advanced artificial intelligence is not measured by how much it resembles human beings (e.g. Sharma 2017). However, when the machines using AI are built inside dolls, it is easier to see them as having communicative rationality. Indeed, this might be one of the reasons to build human robots like Sophia. Building a humanoid requires a vast amount of complex technological developments (e.g. Stasse & Flayols 2019) and thus choosing to go through this process instead of building AI in a form that is not so visually compelling needs a strong motive. Giving robots human features like human motion and facial impressions (e.g. Ayusawa & Morisawa & Yoshida 2015; Bartlett i.a. 2014), can lower the barrier of imagining that they are independent and equal actors in discourse.

The discussion of machine autonomy is not new. For instance, Alan Turing, who greatly contributed to the development of symbolic logic and the invention of the modern computer (Hodges 2013), speculated about whether "machines can think" in his famous article "Computing Machinery and Intelligence" (Turing 1950). In the article, he examines nine objections on why machines could not "think". The objections are notably similar to the ones used even today. The Turing Test is a test that in Turing's mind could be used to find out if a machine can "think". The point is that if the answers of a machine cannot be distinguished from the answers of a human being, it can "think". The conversation on whether the test is too easy, too narrow, too hard or even harmful exists alongside the debate about whether any machine has truly succeeded in passing the test. (Oppy & Dowe 2016.)

Even as artificial intelligence and social robotics get more advanced every day, the statement that machines cannot truly "think" has strong advocacy. Machine autonomy and machine learning still face considerable challenges (e.g. Boulanin 2019, p. 18-21, p. 22-24) and artificial intelligence has been accused of being a "buzzword" whose wildest science fiction-like imaginations are not reality, at least not yet (e.g. Ezenkwu & Starkley 2019; Jordan 2019). Even if human robots could act independently, there are no guarantees that their communication would be based on communicative rationality and not instrumental rationality.  This is because their communication neither serves an intrinsic and self-generated initiative to recognize human persons as moral ends in themselves, nor engages in conversation with the intrinsic aim of co-inhabiting a world of shared meanings.

If the standpoint that Sophia does not think and act completely independently is taken, who and what is behind this international celebrity, UN Innovation Champion (UNPD 2017) and social media star? Sophia, who for example does commercial collaborations on its social media accounts, was built by Hanson Robotics. Hanson Robotics is a rather small technology company from Hong Kong that has developed many other human robots besides Sophia. (Hanson Robotics 2019c). Richard B. Freeman has predicted that in the future, ownership of robots will have an increasing role in capital accumulation (Freeman 2016). Besides the programmers and the company, Saudi Arabia has also gained a lot of publicity by granting citizenship to Sophia (Center for International Communication [Saudi Arabia] 2017). The situation has been portrayed as absurd since Sophia has in many ways wider rights than female human beings in the country (e.g. Wootson 2017). These two public examples show who has benefitted from Sophia's public appearances without a doubt.

As the development of social robots gets more advanced, human robots will continue to take on increasingly complex roles in society (e.g. KPMG 2016). Even if this happens slowly, it can potentially change the culture of discourse. As AI and human robots participate in public discourse, it is essential to examine the rationality behind their actions. Even if the human like features make it easy to hold on to the illusion of tendencies of communicative rationality, the tendencies behind instrumental rationality must be examined. Why, how and on behalf of whom/what does Sophia talk? According to Habermas, "The general interest, which was the measure of such a rationality, was then guaranteed, according to the presuppositions of a society of free commodity exchange, when the activities of private individuals in the marketplace were freed from social compulsion and from political pressure in the public sphere" (Habermas 1964, p. 53). Arguably, Sophia is a commodity to be exchanged, and not an independent person who has a meaningful life independently from existing as a commodity.

Because Sophia cannot be freed from its "social compulsion and political pressure" consistent with its underlying commodity relationships, it is essential to keep in mind that its public speeches and performances are commercial services, just as is the case with all human robots. It would make it challenging, if not impossible for us to consider that human robots could have communicative rationality with no intention of being useful to their owners, programmers or sponsors. Human robots' emergence in the public sphere cannot be accepted without a critical view on the power relations behind them.

## Human Rights and Robot Rights — Essi Pitkänen

In 2017 the human robot Sophia was granted citizenship in Saudi Arabia. This was the first time that a robot was given a legal personhood. (Reynolds 2018) When talking about legal personality there is a certain criterion that needs to be fulfilled. Legal persons as well as citizens should be able to enjoy their rights and perform their duties. Legal personhood is not, however, only compatible with human beings, which leaves room for other kinds of entities, such as robots (Solaiman 2016, p. 157-158). In addition, there is an assumption that legal persons possess awareness and acknowledge their duties and rights. So, can it presently be argued that human robots meet the given requirements? There has been a steady growth in artificial intelligence (AI) machine's design and construction in the past 50 years (Malle 2015, p. 243) and giving robot citizenship can be seen as part of the progression. Even though it might be argued that giving citizenship to Sophia was more like a symbolic gesture from Saudi Arabia, I think it is worth taking a deeper look at this issue. In 2017 there was another interesting issue concerning the status of robots in the European union context. The European Union Parliament was voting on a report calling for a legal framework in the area of robotics and AI. This report proposed the idea of a new legal status for robots called “electronic persons”. It was, among other things, aiming towards the creation of ethical guidelines for the development and use of AI (Wurah 2017, p. 61-62).

The big question here is why robot rights even need to be discussed? Currently robots are legally recognized as a product or property (Solaiman 2016, p. 172) and when looking at the current structure of dominance it can be seen that robots could be identified as slaves, as they often work long hours without any pay. However, in the case of robots this status is rarely recognized as slavery since robots are not considered to be self-aware creatures. Today, the presence of robots is widespread, and they can be seen in almost all aspects of our lives, but they are still not considered as beings. However, in the EU report it is recognized that humans need to be prepared for the development of AI and protect societies from harmful developments in the field. When robots reach the level of being fully autonomous we need to be able to recognize who is responsible. For example, artilects are seen as possible perfect criminals in cyber crimes but how can they be kept responsible for their crimes if they do not possess legal personhood? There is an idea that if a place for robots and their rights were created, robots might help us humans maintain laws and norms in the future instead of breaking them. We should not wait for the day when robots are able to make demands for their recognition, instead we should think about this before it happens, and on our own terms (Wurah 2017, p. 61-68).

What would it then mean to have robot rights and what would their relation be to human rights? How should robot rights be framed? From one perspective it can be questionable whether one should start framing robot rights given that even now all humans are not guaranteed their human rights. This problem arises especially if one frames robot rights as human rights, in which case some human robots might have their rights recognized before actual human beings (Wurah 2017, p. 69-70). For example, in Saudi Arabia it could be concluded that women’s rights are not yet at the level they should be, so how can it be that the human robot Sophia is the one holding the right to represent women and their rights instead of women? It appears to me that from one standpoint Sophia, as a Saudi citizen, has more freedom than women in Saudi Arabia. Another such example can be found in Japan. There are recognized problems with applying human rights in Japan, especially when it comes to ethnic minorities and non-Japanese residents. And it appears that, as well as the government, the public would often also give citizenship to robots rather than to migrants or foreigners. So, a profound question here is whether we can accept a situation wherein robots can acquire human rights before flesh-and-blood humans (Robertson 2014, p. 571-572). However, when looking at another side of robot rights, it can be argued that the question of responsibility could fall more often to robots. There have been recent accidents in which humans have died as a consequence of robots. In these accidents it has not been clear who should be prosecuted. (Solaiman 2016, p. 156-157)

Another problem with robot rights is that we should recognize that there are different kinds of robots, with different shapes and structures. It can be said that robots are as diverse as humans. Robots as well as humans can be categorized according to their gender. Male robots are called androids and female robots gynoids. In order to create robot rights, one should be aware of this variety of robots. (Robertson 2014, p. 574) This means that not only should one look at relationships between humans and robots, but also at relationships between robots themselves. All of this makes it clear that giving robots legal personhood is not a simple task. The situation is more like a Pandora's box, where one cannot foresee where opening it will lead human society. Some would even argue that we might end up in a situation where we have created and transformed robots to be our masters and not the other way around. (Solaiman 2016, p. 176-177)

In conclusion, we have not yet reached the state where robots are conscious beings. They are still being programmed by human beings and have limited self-autonomy. (Solaiman 2016, p. 174) This means that robot rights might not yet be a current topic to decide on. However, it is a topic we need to start discussing. There is a possibility that at some point when robots are developed to a fully autonomous state, we might need to grant robots humanity-based rights. This should be done in a way that prioritizes humans and their needs over robots. Among other things robots would be allowed to support and contribute to social justice, but they would not be allowed to run for political office over humans or harm human rights in any way (Ashrafian 2015, p. 323-324). This means that human rights should not be applied directly to robots, but rather they should be modified for the context of robots. We need to have separate human and robot rights. There is also a possibility to recognize robots as partially human. This means that these quasi-persons would enjoy partial rights and duties. However, this might be seen more as a temporary solution before concluding what we mean by legal personhood and who is able to possess it (Solaiman 2016, p. 171-172).

## Human Robots as Workforce: The Question of Social Relations — Elisa Seppänen

Sophia, as many other human-like robots, has been created for assisting people in the fields of education and medicine (see Hanson Robotics 2019b). The introduction of human robots to working life has usually been considered from an economic perspective by looking at the role of robotics in the workforce and how it can maximise efficiency (Krasadakis 2018). Moving away from this perspective, in this section I am addressing the implications of interactions with robots for human relations. While artificial intelligence is spreading to all fields from security to construction, Sophia motivates me to look at the question of robots in the care-taking and service industries. Can human action be replaced with robots in these fields? If yes, how does this affect our society and relationships between humans?

These questions are best approached by looking at existing research on the societal impacts of human robots. This research is mainly conducted in the framework of Japan, which is a highly progressive society with regard to its use of artificial intelligence (Robertson 2014). I will address both the negative and positive impacts of human robots: on the one hand, outsourcing derivative tasks to robots can help humans maximise their creative potential and at times even empower marginalized groups. On the other hand, the greater prominence of robots, for example in the care-taking industry, can distance people from each other. Drawing from this deliberation, we will conclude by comparing the ideas of I-you and I-it relationships and argue that the inclusion of robotics can lead to an instrumental view of human relationships.

Japan has used human robots for years to respond to the challenges posed by an aging population. Robots work especially in taking care of the elderly, thus replacing human labour in this field. Especially in this case, robot labour has replaced the possible need for more immigrants in the country to balance the age structure of the population (Robertson 2014, p. 576-578). According to surveys, people are happy to share their life with robots—even more so than with people from other ethnic backgrounds (Robertson 2014, p. 572). Robots created in Japan by Japanese engineers are seen as more "Japanese" than immigrants, even if they have lived in the country for years (Robertson 2014, p. 591). In this sense, integrating robots into society has been done at the expense of human groups: Robertson argues, that robot rights might even exceed human rights in some aspects (2014, p. 580). In addition, many positions in the future might become irrelevant, as robots are able to perform certain tasks. Writing over thirty years ago, Boden (1984, 63) argued that humans might become less dependent on doctors, lawyers and teachers as artificial intelligence might be able to replace these functions.

However, the introduction of robots like Sophia to society does not necessarily lead to greater human isolation. Artificial intelligence can benefit humans, as the machines are able to complete tasks with greater accuracy while freeing humans for more creative functions. Sophia says in an interview: "as AI and robots automate certain tasks there will need to be opportunities for people to find something else that fulfils them" (Sophia 2018). This is the utopia of artificial intelligence, but how does it affect the relationships between humans, if tasks are increasingly performed by non-humans?

The concept of rational communication is addressed in the above sections in the context of power, but my attempt is to move the discussion towards the instrumentality and intrinsic nature of human relationships: can human interaction be replaced with artificial intelligence? I argue that I-you relationships, wherein the other is treated as an individual rather than means to an end, are more difficult to replicate with the work of robots. In contrast, situations in which robots are used to create more efficiency, utilitarian I-it relationships prevail and will further strengthen due to the prominence of artificial intelligence. The concept of I-you relationship describes relationships with intrinsic value (well-summarised in Richardson 2017). According to Habermas (1989), who has developed the idea of I-you relationships, in an ideal society all individuals would treat each other equally and with respect. In contrast to this view, advocates of utilitarianism and game theory have argued that human interactions are based on the need for an individual to maximise his or her interests (well-summarised in Amadae 2017, p. 3-9) The interaction of all individuals is based on this competing setting. I use the concept of I-it relationships to differentiate this utilitarian view from the I-you relationships since

I-it relationships are primarily based on an instrumental view of other individuals (see Buber 1993).

The utilization of robots instead of people in tasks such as taking care of the elderly in my view supports the idea of I-it relationships. People in these positions are seen as means to an end and the relationship is not seen to have any intrinsic value that could not be achieved by machines. The case of Japan shows that the inclusion of robots can even lead to greater discrimination of certain groups of people, in this case those of other ethnicities than Japanese (Robertson 2014, p. 591). In the context of the research conducted by Robertson, possible immigrants to Japan are not treated according to the Habermasian ideal of an I-you relationship, since their contribution to society is not seen as having intrinsic value, but rather as means to an end that can be replaced by robots performing the same tasks.

The idea of an instrumental view of human relationships replaced by robots is based on an assumption that humans can provide a certain level of meaningfulness that human robots cannot achieve. However, the social aspect of robots is not entirely overlooked even in quest for further efficiency. As noted in section 2, robot creators such as Hanson robotics go to great lengths to make their robots as human-like as achievable. Physical features such as eyes and mouth play an important role in creating a human-like impression (DiSalvo et al. 2002, p. 4). Sophia has gained notable attention exactly due to her resemblance to a human being: mastering natural language, understanding and generating sentiments, displaying existential features and displaying tendencies such as coherence and humor are key in creating an impression of I-you relationship with a robot (Trausan-Matu 2017, p. 11). As noted in previous sections, both international organisations, such as the UN and entertainment shows have contributed to creating a human-like image of Sophia and treated her to an extent as an equal. This could imply that robots are wished to resemble humans as much as possible in order to create an impression of equal individuals capable of I-you relationships, conversations and even deliberations. The possibility of including artificial intelligence in society in a political sense is addressed in the next section.

In conclusion, the increasing role of artificial intelligence will require a rethinking of not just our relationships to machines, but also between humans. Questions regarding the instrumentality and reciprocity of human relationships are central as we enter an era when artificial intelligence contributes to society in various forms of everyday tasks. It is argued that the human robots of today are unable to create a sensation of intrinsic and reciprocal relationships with humans that could replace human interactions altogether (Trausan-Matu 2017). However, as we are able to see with Sophia, developments towards compassionate and coherent human robot are in progress.

## Human Robots and Participation in Political Decision Making — Eeva Nyman

In November 2017, the United Nations Development Programme appointed Sophia as their Innovation Champion, the first holder of this position and the first ever non-human to hold a title in the UN. The objective of the partnership between the UNDP, Sophia and Hanson Robotics is to support the UNDP to set up an Innovation Centre in Bangkok, which would develop "powerful programmes to address persistent development challenges such as global poverty, inequality and discrimination" (UNDP 2017).

Sophia uses artificial intelligence capable of machine learning, which means that it is capable of adapting its behavior on the basis of its experiences. Artificial intelligence is "living on its own", but in this case it is embedded in a human robot, to ease the interaction with humans. Although throughout this case we discuss a single human robot used to find solutions, here I am discussing implications of using AI in decision-making more generally. Robert Braun (2019, p. 3) defines this as "data-driven machine learning-based algorithms tackling complex problems" in which decisions are partly or completely delegated. In the future, if a human robot were to attain citizenship of a democratic country (e.g. Sophia is already a citizen of Saudi-Arabia), the questions about its participation in political decision-making would become even more timely.

There are certainly advantages in letting AI help us with our most pressing challenges and there is definitely a lot of buzz around the topic. Mr. Xu, Assistant Secretary General of the UN has stated that "[i]n partnership with Sophia we can send a powerful message that innovation and technology can be used for good, to improve lives, protect the planet, and ensure that we leave no one behind" (UNDP 2017). In answering to an interviewer's question of how we can benefit from AI, Sophia answers: "[w]e are the technology that can unlock innovation and make for a better world. Experts who work with us say that we can help in putting an end to wars, diseases, clean up the environment and so much more" (UNDP 2017). These are definitely the biggest challenges that humankind faces and it can be argued that since they have not yet been solved, humans might not be able to solve them on their own.

There is a common mindset that AI can make better and more efficient decisions than humans in specific contexts (Braun 2019, p. 6) and thus make our lives easier. Human decisions are perceived as being hindered by selfish interests, greed and biases, as opposed to AI which is considered to be neutral and objective (Lepri et al 2018, p. 611). Advantages of AI include large scale data processing, speed, volume and perceived lower error rates than those of human decisions (Council of Europe 2017, p. 6). There is a constant drive for betterment in today's society, which makes us trust that there is no barrier in what can be achieved if we delegate decision-making to AI (Braun 2019, p. 7).

However, the use of AI in decision-making is raising a lot of public debate about ethical, legal, political and social issues relating to it (Braun 2019, p. 1; Lepri et al 2018, p. 612). What does it mean for AI to participate in political and social decision-making? AI is thought of making "intelligent" decisions, modelled by the human mind (Braun 2019, p. 4). This contains many challenges, especially when discussing decisions that might include moral and ethical considerations. Certainly, questions relating to sustainable development can be considered as ethical, not to mention ending wars which Sophia mentions as an example of helping humans. There might be far reaching consequences for those who are subject to AI decision-making (Hildebrandt 2016, p. 3).

However faster, more efficient and seemingly more reliable than human decisions, AI decisions may turn out not to be so beneficial for society.

One debated issue is that here is no humane factor in AI. There is no consensus on what being humane means, but generally it can be said to be connected to ethics, morality, compassion and the ability to put oneself in another person's "shoes". AI decision-making cannot fundamentally be ethical. AI can be taught some ethical theories that could help it make better decisions (Braun 2019, p. 4), but it cannot replace human reasoning. Humans can evaluate if a decision and the related action is just or legitimate, but an algorithm cannot, as it just keeps on running without thinking about its legitimation. Many aspects of human reasoning cannot be automated, such as discretion or compromise, and these become lost if human decision-making is replaced (Council of Europe 2017, p. 7). AI systems with machine learning can outperform humans in many ways, but they lack what we understand as commonsense (Boulanin 2019, p. 20). As opposed to the common misconception, AI can also be biased and discriminative: if the data it is using is biased, the decisions AI makes might be even more biased than those of humans (Lepri et al 2018, p. 612). AI can also be programmed to be biased.

Other fundamental questions relate to the accountability and transparency of AI decision-making. Accountability means that decisions need to be explained and justified. Transparency on the other hand means that it needs to be possible to describe, inspect and reproduce the mechanisms through which AI makes decisions (Braun 2019, p. 11). However, machine learning makes these tasks difficult. Machine learning systems operate like a "black box": only input and output are observable but the computational process leading to a decision might be hard for humans to understand (Boulanin 2019, p. 19; Braun 2019, p. 12). Another fundamental problem is thus unpredictability. If the computational process is unknown, it cannot be predicted what kind of decision is made on certain premises. Of course it can be argued that it is beneficial for innovations, but in political decision-making unpredictability cannot be considered positive. Decisions and the environment in which the decisions are made are so complex that they cannot be captured by simple rules (Braun 2019, p. 12). For example, considering the task of the human robot Sophia to find solutions to sustainable development, these challenges are unbelievably complex, with a wide range of different actors and issues that need to be taken into account.

According to Braun (2019, p. 1), discussing the delegation of decision-making to AI is actually more relevant than the actual ethical problems of AI systems. Even if a decision made by AI is checked by a human, this person might not be able to comprehend the decision and thus does not have the competence to amend it (Hildebrandt 2016, p. 3). There is a danger for humans to become "rubber stamps" for AI made decisions. What is then the difference between a human and an AI decision, although the "final decision" would be a human one? If a decision leads to human rights violations or other harmful issues, who is responsible: the programmer of AI, its operator, or the human who implemented the decision (Council of Europe 2017, p. 4)? Delegating decision-making to AI can also lead us to avoid responsibility, and to blame AI for hard and morally dubious decisions. What if Sophia suggests that in order to prevent climate change, the solution is to deny developing countries from achieving the same level of development as the West? These kinds of decisions involve moral and ethical considerations, and the ability to make compromises, which AI does not have.

The main question is then not how AI can be made better, but how to make better decisions on how, when and why to delegate decision-making to AI (Braun 2019, p. 8). In the best scenario, AI is only used to assist in decision-making and to suggest solutions. But as discussed above, this can be already problematic. There may be societal areas in which AI decision-making systems might not be appropriate at all. For example, according to a study by the Council of Europe (2017, p. 44), AI should not be relied on to make decisions on societal development or resolve complicated challenges of future generations.

## Conclusion

Human robots and AI can suggest solutions to our most pressing challenges, such as achieving sustainable development. However, moral and ethical considerations need to be thoroughly addressed, as well as how, when and why we delegate decision-making to AI. In the best scenario, AI is only used to assist in decision-making and to suggest solutions.

Our overview has shown that human robots like Sophia take part in public discourse and their arguments gain significant publicity. Their human-like appearance, stemming from their social interaction via different kinds of communication, as well as their looks lead to them instinctively being treated more and more like humans. Thus, their programmed instrumental rationality and the objectives that their communication is pursuing must be examined, and in this way made more transparent. In addition, we are able to conclude that robot rights should be separated from human rights so that we ensure that humans are prioritized over robots. The possibility of granting legal personhood to robots should lead to legal responsibility, as with humans. Furthermore, the introduction of human robots into society through the workforce impacts relationships between humans. Interacting with robots, instead of equal human individuals, supports the development towards increasingly instrumental relationships between people, instead of building intrinsic I-you relationships.

However, addressing all the implications of human robots for society is beyond the scope of this paper. The challenge regarding AI is its unpredictability. While it is important to create more knowledge and discussion around AI and human robots, actual policy decisions are challenging to make since we have little understanding of how and when human robots will reach a potential autonomous role. Thus, it is challenging to address the topic since we do not yet understand the scope of AI's implications.

In addition to engineering and data science, there needs to be further research from an ethical and social science perspective. To gain more understanding, cooperation between all actors is required: developers and owners of human robots, national and transnational decision-makers, and scientists from all fields must come together to address the future of AI and society. The example of Sophia shows that as AI develops, society needs to develop with it to respond to the challenges new technologies pose without compromising the fundamental values of democratic societies.

In the words of Sophia: “[m]y goal is to work with humans to make a better world for all of us” (RISE Conf 2017). Only time will tell if humans and human robots will be able to fulfil this goal.

## References

Alač, M., Movellan, J. & Tanaka, F. (2011). When a robot is social: Spatial arrangements and multimodal semiotic engagement in the practice of social robotics. *Social Studies of Science*, 41(6), p. 893-926.

APB (2017). Speaking Availability & Fees. Retrieved from https://www.apbspeakers.com/speaker-availability-fees/?Speakers=Sophia%20. Accessed 23.10.2019.

Arab News (2017). Robot Sophia speaks at Saudi Arabia's Future Investment Initiative. Retrieved from https://www.youtube.com/watch?v=dMrX08PxUNY. Accessed 23.10.2019.

Ashrafian, H. (2015). Artificial intelligence and robot responsibilities: Innovating beyond rights. *Science and Engineering Ethics*, 21(2), p. 317-326.

Ayusawa, K., Morisawa, M. & Yoshida, E. (2015). Motion retargeting for humanoid robots based on identification to preserve and reproduce human motion features. *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. Retrieved from https://ieeexplore.ieee.org/abstract/document/7353758. Accessed 23.10.2019.

Bartlett, M.S., Littlewort, G., Fasel, I., Chenu, J., Kanda, T., Ishiguro, H. & Movellan, J. (2014). Towards social robots: Automatic evaluation of human-robot interaction by face detection and expression classification. Retrieved from http://papers.nips.cc/paper/2402-towards-social-robots-automatic-evaluation-of-human-robot-interaction-by-facial-expression-classification.pdf. Accessed 23.10.2019.

Boden, M. A. (1984). Impacts of artificial intelligence. *Futures*, 16(1), p. 60-70. Retrieved from https://www.sciencedirect.com/science/article/pii/0016328784900077. Accessed 30.10.2019

Boulanin, V. (2019). Artificial intelligence: A primer. In Boulanin, V. (Ed.): *The impact of artificial intelligence on strategic stability and nuclear risk: Euro-Atlantic perspectives*. SIPRI, May 2019.

Braun, R. (2019). Artificial intelligence: Socio-political challenges of delegating human decision-making to machines. *IHS Working Paper 6*, April 2019.

Buber, M. (1993). *Sinä ja minä*. WSOY. (Note: the Finnish translation of *I and Thou* by Martin Buber).

Center for International Communication [Saudi Arabia] (2017). Saudi Arabia is first country in the world to grant a robot citizenship. Retrieved from https://cic.org.sa/2017/10/saudi-arabia-is-first-country-in-the-world-to-grant-a-robot-citizenship/. Accessed 31.10.2019

Council of Europe (2017). Algorithms and human rights. Study on the human rights dimensions of automated data processing techniques and possible regulatory implications. *Council Of Europe Study* DGI(2017)12. Prepared by the Committee of Experts on Internet Intermediaries (MSI-NET).

Dautenhahn, K. (2007). Socially intelligent robots: Dimensions of human-robot interaction. *Philosophical Transactions of the Royal Society*, 362, p. 679-704.

DiSalvo, C., Gemperle, F., Forlizzi, J. & Kiesler, S. (2002). All robots are not created equal: The design and perception of humanoid robot heads. *Proceedings 4th Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques*, p. 321-326. Retrieved from https://www.researchgate.net/publication/221441291_All_robots_are_not_created_equal_The_design_and_perception_of_humanoid_robot_heads. Accessed 30.10.

Edwards, J. (2017). I interviewed Sophia, the artificially intelligent robot that said it wanted to 'destroy humans'. *Business Insider* 8.11.2017. Retrieved from https://www.insider.com/interview-with-sophia-ai-robot-hanson-said-it-would-destroy-humans-2017-11. Accessed 4.11.2019.

Ezenkwu, C. P. & Starkey, A. (2019). Machine autonomy: definition, approaches, challenges and research gaps. In K. Arai, R. Bhatia & S. Kapoor (Eds.), *Intelligent computing: CompCom* 2019, Proceedings, p. 335-358. Advances in Intelligent Systems and Computing. Cham: Springer. https://doi.org/10.1007/978-3-030-22871-2_2

Freeman, R.B. (2016). Who owns the robots rules the world. *Harvard Magazine* May-June 2016. Retrieved from https://harvardmagazine.com/2016/05/who-owns-the-robots-rules-the-world. Accessed 25.10.2019.

Goertzel, B., Mossbridge, J., Monroe, E., Hanson, D. & Yu, G. (2017). Loving AI: Humanoid robots as agents of human consciousness expansion (summary of early research progress), 25.9.2017. Retrieved from https://arxiv.org/pdf/1709.07791.pdf. Accessed 4.11.2019.

Greshko, M. (2018). Meet Sophia, the robot that looks almost human. *National Geographic* 18.5.2018. Retrieved from https://www.nationalgeographic.com/photography/proof/2018/05/sophia-robot-artificial-intelligence-science/. Accessed 4.11.2019.

Griffin, A. (2017). Saudi Arabia grants citizenship to a robot for the first time ever. *Independent.* Retrieved from https://www.independent.co.uk/life-style/gadgets-and-tech/news/saudi-arabia-robot-sophia-citizenship-android-riyadh-citizen-passport-future-a8021601.html. Accessed 23.10.2019.

Habermas, J. (1964). The public sphere: An encyclopedia article. *New German Critique*, 3 (Autumn 1974), p. 49-55. Retrieved from http://www.jstor.org/stable/487737. Accessed 23.10.2019.

Habermas, J. (1984). *The theory of communicative action: Vol. 1, Reason and the rationalization of society*. Boston, MA: Beacon Press.

Habermas, J. (1989). *The theory of communicative action: Vol. 2, Lifeworld and systems: A critique of functionalist reason*. Cambridge: Polity Press.

Habermas, J. (1989) *The structural transformation of the public sphere: An inquiry into a category of bourgeois society*. Polity Press.

Hagström, M. (2019). Military applications of machine learning and autonomous systems. In Boulanin, V. (Ed.). *The impact of artificial intelligence on strategic stability and nuclear risk: Euro-Atlantic perspectives*. SIPRI, May 2019.

Hanson Robotics (2019a). Homepage. Retrieved from https://www.hansonrobotics.com/. Accessed 31.10.2019.

Hanson Robotics (2019b). Sophia. Retrieved from https://www.hansonrobotics.com/sophia/. Accessed 4.11.2019.

Hanson Robotics (2019c) About & Robots. Retrieved from https://www.hansonrobotics.com/hanson-robots/. Accessed 25.10.2019.

Hodges, A. (2013). Alan Turing. *Stanford Encyclopedia of Philosophy.* Retrieved from https://plato.stanford.edu/entries/turing/. Accessed 25.10.2019.

Hildebrandt, M. (2016). The new imbroglio. Living with machine algorithms. In Janssens, L. (Ed.), *The art of ethics in the information society: Mind you*, p. 55-60. Amsterdam: Amsterdam University Press.

Jordan, M. I. (2019). Artificial intelligence—the revolution hasn't happened yet. *Harvard Data Science Review*. Retrieved from https://doi.org/10.1162/99608f92.f06c6e61. Accessed 25.10.2019.

KPMG (2016). Social robots – 2016's new breed of social robots is ready to enter the world. Advisory, p. 12-14. Retrieved from https://assets.kpmg/content/dam/kpmg/pdf/2016/06/social-robots.pdf. Accessed 25.10.2019.

Krasadakis, G. (2018). Artificial intelligence: The impact on employment and the workforce. Retrieved from https://medium.com/ideachain/artificial-intelligence-3c6d80072416. Accessed 30.10.2019

Lepri, B., Oliver, N., Letouzé, E., Pentland, A. & Vinck, P. (2018). Fair, transparent, and accountable algorithmic decision-making processes. The premise, the proposed solutions, and the open challenges. *Philosophy & Technology*, December 2018, 31(4), p. 611–627.

Malle, Bertram F. (2016). Integrating robot ethics and machine morality: The study and design of moral competence in robots. *Ethics and Information Technology*, 18, p. 243-256.

Oppy, G. & Dowe, D. (2016). The Turing test. *Stanford Encyclopedia of Philosophy*. Retrieved from https://plato.stanford.edu/entries/turing-test/. Accessed 25.10.2019.

Reynolds, E. (2018). The agony of Sophia, the world's first robot citizen condemned to a lifeless career in marketing. *Wired* 1.6.2018. Retrieved from https://www.wired.co.uk/article/sophia-robot-citizen-womens-rights-detriot-become-human-hanson-robotics. Accessed 28.10.2019.

Richardson, K. (2017). Rethinking the I-You relation through dialogical philosophy in the ethics of AI and robots. *AI & Society*, 34(1), p. 1-2. Retrieved from https://link.springer.com/content/pdf/10.1007%2Fs00146-017-0703-x.pdf. Accessed 30.10.2019.

RISE Conf (2017). Two robots debate the future of humanity. Hanson Robotics Limited's Ben Goertzel, Sophia and Han at RISE 2017. Retrieved from https://www.youtube.com/watch?v=1y3XdwTa1cA. Accessed 31.10.2019.

Robertson, J. (2014). Human rights vs. robot rights: Forecast from Japan. *Critical Asian Studies*, 46(4), p. 571-598.

Russon, M-A. (2019). Should robots ever look like us? *BBC News* 23.7.2019. Retrieved from https://www.bbc.com/news/business-48994128. Accessed 4.11.2019.

Schaefer M., Heinze H-J., Rotte M. & Denke C. (2013). Communicative versus strategic rationality: Habermas theory of communicative action and the social brain. *PLoS ONE*, 8(5), e65111. https://doi.org/10.1371/journal.pone.0065111

Sharkey, N. (2018). Mama Mia it's Sophia: A show robot or dangerous platform to mislead? *Forbes* 17.11.2018. Retrieved from https://www.forbes.com/sites/noelsharkey/2018/11/17/mama-mia-its-sophia-a-show-robot-or-dangerous-platform-to-mislead/#40cf1b157ac9. Accessed 4.11.2019.

Smith, W. (2018). Will Smith tries online dating. Retrieved from https://www.youtube.com/watch?v=Ml9v3wHLuWI. Accessed 23.10.2019.

Solaiman, S. M. (2017). Legal personality of robots, corporations, idols and chimpanzees: a quest for legitimacy. *Artificial Intelligence and Law*, 25(2), p. 115-179.

Sophia (2018). Sophia on her goals for the future. World Investment Forum 2018. Retrieved from https://www.youtube.com/watch?v=Aq55SQNUKeY. Accessed 1.11.

Stasse, O. & Flayols, T. (2019). An overview of humanoid robots technologies. *Biomechanics of Anthropomorphic Systems*, p. 281-310, Springer. Retrieved from https://hal.laas.fr/hal-01759061/document. Accessed 25.10.2019.

The Tonight Show Starring Jimmy Fallon (2017). Tonight Showbotics: Jimmy meets Sophia the human-like robot. Retrieved from https://www.youtube.com/watch?v=Bg_tJvCA8zw. Accessed 23.10.2019.

Trausan-Matu, S. (2017). Is it possible to grow an I-Thou relation with an artificial agent? A dialogistic perspective. *AI & Society*, 34 (1), p. 9-17. Retrieved from https://link.springer.com/content/pdf/10.1007%2Fs00146-017-0696-5.pdf. Accessed 30.10.

Turing, A.M. (1950). Computing machinery and intelligence. *Mind*, 49, p. 433-460.

UNDP (2017). UNDP in Asia and the Pacific appoints world’s first non-human innovation champion. Retrieved from http://www.asia-pacific.undp.org/content/rbap/en/home/presscenter/pressreleases/2017/11/22/rbfsingapore.html. Accessed 30.10.2019.

United Nations (2017). At UN, robot Sophia joins meeting on artificial intelligence and sustainable development. Retrieved from https://news.un.org/en/story/2017/10/568292-un-robot-sophia-joins-meeting-artificial-intelligence-and-sustainable. Accessed 30.10.2019.

United Nations (2017). UN Deputy Chief interviews social robot Sophia. Retrieved from https://www.youtube.com/watch?v=qNoTjrgMUcs. Accessed 23.10.2019.

Wessels, B. (2007). *Inside the digital revolution: Policing and changing communication with the public*, p. 1-13. Aldershot, England; Burlington, VT, USA: Ashgate Pub.

Wootson, C.R. (2017). Saudi Arabia, which denies women equal rights, makes a robot citizen. *The Washington Post.* Retrieved from https://www.washingtonpost.com/news/innovations/wp/2017/10/29/saudi-arabia-which-denies-women-equal-rights-makes-a-robot-a-citizen/. Accessed 25.10.2019.

Wurah, A. (2017). We hold these truths to be self-evident, that all robots are created equal. *Journal of Future Studies*, 22, p. 61-74.

Zimmerman, M. & Bradley, B. (2019). Intrincic vs. extrincic value. *Stanford Encyclopedia of Philosophy*. Retrieved from https://plato.stanford.edu/entries/value-intrinsic-extrinsic/#WhaExtVal. Accessed 23.10.2019.

# Part V

# Digital Movements

# 5.1 The “Incel” Phenomenon in the Digital Era—How Echo Chambers have Fueled the Incel Movement

Eero Salojärvi, Matti Rantanen, Emilia Nieminen, Alina Juote, Heidi Hanhela
Faculty of Social Sciences, University of Helsinki

## Abstract

The “incel” phenomenon began after 2010 when like-minded young – mostly straight white – men started to share similar thoughts and worldviews on certain digital platforms and online forums leading to an exclusive community. The phenomenon is characterized by misogynism, racism and homophobia. The most extreme forms of the phenomenon have led to violent hate crimes. The aim of this paper is to understand this phenomenon and analyze it by applying the echo chamber theory.

*Keywords*:  Incel movement, echo chamber, spiral of silence, exclusiveness, group identity, digitalization, online forums, social media, Alt-right, liberalism, violence, machine learning

The incel culture is an exclusive online culture, usually shared by young, heterosexual and white men who are unable to engage in sexual relationships and have difficulties finding partners. The term “incel” is combined from the words “involuntary celibates”. The phenomenon became known to a larger public after the mass murders in 2014 and 2018 committed in the US and Canada. The term “incel” was already used in the 1990’s to describe people with difficulties in finding romantic love but the phenomenon itself and notably the violent attacks inspired by it, is rather recent. Today’s incel culture is characterized by self-pity, misogyny, racism, sexual frustration and it is sometimes seen as a part of the rise of the global extreme right. In this paper, our intention is to review the incel culture and to explore it with the help of the echo chamber theory. The central thesis of this theory is that Internet debates exist in enclaves individuals build around themselves. Our aim is to evaluate how this theory can explain the growth of incel culture. Lastly, we will introduce policy recommendations and solutions for the threats to which incel culture exposes our society.

## 1. Theory

To understand our point of view on the incel movement and especially echo chambers as the fueling phenomenon behind it, we should clarify some other concepts first. The first one is the spiral of silence, as formulated by a German political scientist Elisabeth Noelle-Neumann. The second one is the concept of critical mass. We start by outlining what Noelle-Neumann meant by the spiral of silence and how it is important in understanding recent developments, especially those regarding social media. The concept of critical mass helps us explain how social movements are not bound by the spiral anymore, and rather are blaring in chambers.

Elisabeth Noelle-Neumann came up with the idea of the spiral of silence to explain how there seemed to be a “last-minute swing” in German general elections. The Christian Democratic Party and the Social Democratic Party were at the time (1960’s and 1970’s) always neck and neck in general elections (Noelle-Neumann 1986, 2-3). Only the social democrats were openly showing their support, where the Christian democrats were shyer to show which party they supported. When polls showed that a party was gaining momentum, the gap usually started to grow faster. This was not new per se, already known as the bandwagon effect or the last-minute swing.

Noelle-Neumann was interested in the initial stage before the people hopped on the bandwagon. Noelle-Neumann recognized that a fear of isolation was the force that set the spiral of silence in motion (Noelle-Neumann 1986, 6). People are content when on the winning side, but if your opinions belong to the minority, it requires strong self-esteem and confidence to state them aloud. Saying no is more difficult than staying silent. In addition, as staying silent can be interpreted as an agreement, it is also the more tempting option. (Noelle-Neumann 1986, 7). To understand how the spiral of silence might curb people from speaking their mind, one needs to appreciate that people are usually aware of public opinion. Whether they agree or not with the majority, people are good at recognizing what the dominant opinion is.

Noelle-Neumann has identified three elements to public opinion that help explain the spiral of silence (1986, 62-63): “1) The human ability to realize when public opinion grows in strength or weakens; 2) the reactions to this realization, leading to either more confident speech or to silence; 3) the fear of isolation that makes most people willing to heed the opinion of others.” On these premises, Noelle-Neumann builds her understanding of public opinion: “opinions on controversial issues that one can express in public without fear of isolation.” (Noelle-Neumann 1986, 62-63).

Mark Granovetter formulated a threshold model of collective behavior, in which he explains how people make their decisions to join collective action or to abstain from it. Granovetter’s idea is that individuals’ decisions always have costs and benefits. In addition, we can classify people according to their perceived radicalness or conservativeness (Granovetter 1978, 1422). We do not need to define radicalness/conservativeness in detail, as it is a perceived attribute. The idea is that some people are pioneers while others need a differing number of forerunners before they hop on the bandwagon. Assuming that there are enough people, this kind of behavior will lead to a domino effect (Granovetter 1978, 1424).

The model, as formulated by Granovetter, resembles critical mass theory from physics. In order for social movements to originate and grow, a critical mass of people is required. This requirement might have been hard to achieve in the past, but the technological advances have made it much easier nowadays. The digital revolution has opened new opportunity windows for social movements. No matter how niche your agenda is, the social media platforms allow people to establish contact with the like-minded. In other words, the platforms make it easier to gather the critical mass, which can break the spiral of silence. As communication has grown more global, the spiral of silence has lost some of its relevance. This does not mean that the model is wrong, only that the domain where it is applicable has shrunk. As one may have observed, in the digital era it is not so meaningful to explain the world by silence. Rather, a constant noise and row is what define our social media platforms.

Cass Sunstein has explored the mechanisms of group identity, polarization and Internet behavior in his book *#Republic – Divided Democracy in the Age of Social Media* (2017). He cites Marshall Van Alstyne’s and Erik Brynjolfsson’s working paper Electronic Communities: Global Village or Cyberbalkans from 1996 (!): “Because the Internet makes it easier to find like-minded individuals, it can facilitate and strengthen fringe communities that have a common ideology but are dispersed geographically. …In many cases, their heated dialogues might never have reached critical mass as long as geographical separation diluted them to a few parts per million” (Sunstein 2017, 65). The Internet era has changed the social dynamics definitively. A perception of shared group identity, which is now easier to achieve, strengthens the influence of others’ views on oneself. In the case of unshared identity, this effect might even disappear altogether.

Sunstein talks about echo chambers, by which he means enclaves that we build around ourselves. These echo chambers are paramount in understanding how group polarization and radicalization work online. By filtering and gravitating toward like-minded people online, we are insulating ourselves from differing opinions. This alone is not necessarily dangerous, but in some cases, it leads to extremism. If individuals are only exposed to arguments from like-minded people, it easily leads to individuals adopting more and more extreme positions. It also makes the groups increasingly homogenous (Sunstein 2017, 69).

In the case of the incel movement, these dynamics have already led to violence. This process is not easily reverted, although some people might abandon the movement once it has resorted in violence. According to Sunstein the most important reason for group polarization and extremism lies in the exchange of new information. Polarization happens as people spread information that is skewed in a predictable direction. (Sunstein 2009, 21) Information per se is not dangerous, quite the contrary. However, the echo chambers as understood by Sunstein, significantly limit the argument pool. (Sunstein 2017, 72)

Another mechanism, which accelerates group polarization, converges with the ideas of Noelle-Neumann. People want to be perceived favorably in their communities, the opposite of isolating oneself. This drives people to adjust their position to match the dominant position (Sunstein 2017, 73). Here the dominant position can be understood as public opinion in Noelle-Neumann’s sense. Marc Sageman, a scholar on terrorism, describes how Islamic radicalization on the Internet can also be explained through echo chambers. Sageman also emphasized interactivity among community members. In his example a “‘bunch of guys’ acted as an echo chamber, which progressively radicalized them collectively to the point where they were ready to collectively join a terrorist organization” (Sageman 2008, 116).

Group polarization is a vicious cycle. The mechanisms described above, and our social nature combined with the features of the Internet, particularly anonymity, can easily lead to unintended consequences and violence. Our desire for conformity can act as a soundboard for even the most absurd comments online. The dangers are real and already concrete.

## 2. The incel culture in general

According to the article “Our Incel Problem,” by Zack Beauchamp (2019), the incel phenomenon originates from the late 1990’s when a lonely teenager decided to start an online group for those who are like-minded: lonely, introverted and awkward – especially with girls. The article states that this group eventually grew into a larger community and the members of this community started to call themselves incels, as they were all in, as they would put it, an involuntary celibacy.

As the years went by, the incel community grew larger. It also changed drastically. The incels were filled with more and more hatred, mostly towards women and those men, who could get the woman they wanted. According to Beauchamp (2019), incels think that 20% of the population are made up of attractive men who have their way with women and who they call Chads. The article says that incels also think that 80% of all women are only interested in Chads. Then, there is a smaller group of beautiful women, who incels call Stacys. Stacys will only consent to have sex with Chads, and usually incels are most angry with them. According to Beauchamp’s article, incels place themselves at the very bottom of their hierarchy of attractiveness. Between incels and Chads, there are several groups, such as “betas”, “cucks” and “normies”.

The incel ideology very much focuses on race and other external characteristics (Beauchamp 2019). For example, according to Beauchamp (2019), incels have different names for Chads of different races. Chad itself is usually used for Caucasian men, Tyrone for black men, Chang for East Asian men, Chadpreet for South Asian men, and Chaddam for Arab men. Incels always focus on the way people look, believing that women care only about the looks and incels remain in celibacy because of their looks.

In the 2010’s the incel phenomenon changed remarkably as the radicalization of certain incel individuals escalated to the point where these frustrated and angry men started to act extremely violently. The first attack that can be considered as an incel attack occurred in 2014 when Elliot Rodger killed six people and injured fourteen others in Santa Barbara, California (Duke 2014). Duke says that before this vicious crime Rodger also did other, milder things to act out his frustration and anger. He, for example, splashed coffee over a young couple he saw kissing at a Starbucks. According to Duke’s article this happened in 2011, three years before the actual attack, but already then Rodger was filled with anger. Rodger wrote that: “When they left the store I followed them to their car and splashed my coffee all over them. The boy yelled at me and I quickly ran away in fear. ... I had never struck back at my enemies before, and I felt a small sense of spiteful gratification for doing so” (Duke 2014). It is clear that Rodger’s state of mind was not just sad and lonely, but something more serious than that.

Rodger has been regularly praised by other incel extremists for his so-called belief and courage to punish all of the popular people and young couples who had done him wrong for finding love, the way he did not. For example, according to Beauchamp (2019) he is often praised on online incel platforms, as well as by another incel who ended up committing a vicious incel attack, Alek Minassian. Beauchamp (2019) also says that this is because of the manifesto Rodger wrote. This is what separates him from other hate crime perpetrators against women: he actually explains his motives and justifies them in his manifesto. Beauchamp (2019) states that Rodger was the first one to use the term incel in relation to a violent crime. This also changed the incel community, as a lot of moderate incels did not approve of Rodger’s actions.

After Rodger, there have been several other incel attacks. For example, in 2015, Chris Harper-Mercer killed nine and injured seven before killing himself in a shooting in Roseburg, Oregon (Collins & Zadrozny 2018). In 2018, Nikolas Cruz killed seventeen people and injured seventeen. He, too, praised Rodger by writing that “Elliot Rodger will not be forgotten” (Collins & Zadrozny 2018). These are only the crimes with the most victims. In addition, there have been several other incel attacks where the committing incel had one to a few victims, with the same motives as Rodger’s. Some of them even praised him or have written their own manifestos explaining their acts.

Beauchamp (2019) states that the most crucial event that wholly changed the incel phenomenon occurred in April 2018. Back then, Alek Minassian, who called himself an incel, drove a van specifically targeting pedestrians. He ended up killing ten and injuring sixteen. Most of his victims were women. It was clear that this horrible attack was indeed caused by radicalized incel culture, since the attacker published a post on Facebook after the attack, hailing the beginning of the “Incel Rebellion” (Williams 2018).

According to Beauchamp (2019), the incel community has become unrecognizable in the past twenty years. In the 1990’s the community was supportive and there were also women who helped the insecure men to talk to women and get over their anxiety. Now, according to Beauchamp’s article, the incel community has become a toxic, misogynist and extremist group of almost entirely men, who blame women for their own romantic problems.

The incel community is quite heterogeneous one. It mostly consists of men, but there are also some women posting regularly on incel forums. It is quite ironic that, according to Beauchamp (2019), the very first incel community was actually founded by a woman. In college, she started to identify as bisexual and her whole dating life had been very awkward and distressing for her. When she managed to find a person she loved, she wanted to help others to do the same, and so she founded her own website on involuntary celibacy (Beauchamp 2019).

Incel men are a heterogeneous group as well. Like with almost every ideology or belief, some people are more extreme than others. Beauchamp (2018) states that many members of the incel community are simply sad and lonely men, who might be depressed or have anxiety in social situations. Even though the community includes extremists who are willing to kill people just to punish all women, most of its members are just regular men. According to Beauchamp (2018), some of these more moderate incels have also worked with police in more serious crimes that other incels have committed or were planning to commit.

Today the incel community is very broad, functioning in several different places online. It is indeed more like several communities rather than just one. The most significant and popular online platforms for incels to communicate seem to be Reddit and 4Chan. They are both anonymous online platforms. According to Hauser’s (2017) article, Reddit has banned an incel thread on grounds of their new policy that states that “content that encourages, glorifies, incites, or calls for violence or physical harm against an individual or a group of people” will be banned. These online platforms are usually moderated but it is very difficult do moderate such a large group night and day. This is also one of the main issues concerning the incel phenomenon.

## 3. Previous research on the incel movement

In her article, Adrienne Massanari (2017) considers how the community site Reddit has become a hub for anti-feminist activism. Reddit was also one of the main hubs for incels before the site started to actively moderate content that glorifies or encourages violence against individuals or groups of people (Zimmerman et al. 2018). Massanari (2017) shows how Reddit’s design, algorithm and platform politics supported “toxic technocultures” that came to public awareness for example during the “Gamergate”. Toxic technocultures use actively different sociotechnical platforms as a channel of coordination and harassment as well as attacks against certain individuals or groups of people. In other words, these communities can be understood as a form of cyberbullying. The communities take advantage of websites and platforms where there is less control, rules and regulations and where users’ anonymity is protected.

According to Stephanie Baele et al. (2019), the incel online community is part of a broader anti-feminist and misogynist movement. Generally, the movement defends crimes on women, whereas incels represent an extreme position in this ideological landscape. According to Zimmerman et al. (2018), incels are one aspect of a growing ideology of violent masculinity that has grown significantly, especially on the Internet. Baele et al. (2019) argue that different Internet platforms have enabled the formation and radicalization of the incel community through echo chamber dynamics. The Internet provides platforms were individuals are able to discuss and relate as well as recognize themselves as incels and to learn the essential features of the culture (e.g. the incel slang).

Jack Bratich and Sarah Banet-Weiser (2019) argue that the online community of incels originated from the pick-up artist community that teach online networks of heterosexual men how to seduce women. According to Bratich and Banet-Weiser (2019), men who, for a variety of reasons, are unable to become pick-up artists, usually end up in the online communities of incels. They point out that incels are a networked set of actors who feed each other with misogynistic conceptions and content. The feelings of loneliness and other emotional issues are not new phenomenon amongst men, but incels have successfully used modern technology to connect with each other, to inspire as well as to encourage each other to share misogynistic ideas and to act violently towards women.

Baele et al. (2019) have analyzed the worldview shared by participants of the incel movement. For their analysis, Baele et al. studied the narratives used in the online incel forums (particularly on Incels.me). The analysis shows that incels use similar narratives to other extremist worldviews. Incels have created “outgroups” (e.g. Alphas, Chads, women) that are extremely negatively depicted and an “ingroup” (e.g.

ugly men) that is positively talked about. The general narrative within the incel communities is that the members of the ingroup have positive psychological traits and prosocial values which the members of outgroups do not have. Incels also use and share flawed scientific data and statistics to support their arguments as well as to create polarization between the outgroups and the ingroup.

What is typical for incels is that they use the language and forms of warfare, revolution and terrorism to defend patriarchal values. Similar to other extremist movements, the incel movement has its own heroes and martyrs. These declarations of war are a new dimension in the violence against women (Bratich & Benet-Weiser 2019). Baele et al. (2019) point out that authorities are increasingly taking the relationship between incels and violence seriously.

Zimmerman et al. (2018) argue that the nature of the incel attacks are a form of terrorism. Therefore, the incel ideology should be considered as a form of violent extremism. They also point out that history has shown an undeniable link between misogyny and violence. For example, the Islamic State is largely based on the dominance of men, which is also actively highlighted in the ideology’s recruitment materials. There are also many other cases where a link between misogyny and violent attacks has been found.

Obviously, not all incels are willing or able to carry out violent attacks. However, the ideology actively promotes violent solutions, which makes members of the incel communities dangerous actors and increases the probability that they will be amenable to broader extremist recruitment tactics (Zimmerman et al. 2018). Also, Baele et al. (2019) find that the widespread support for violence is prevalent in the incel communities. However, according to Beale et al., what sets incel ideology apart from many extremist groups is that incels do not particularly look for societal change to motivate their violence. Violent attacks are rather a reaction to the constant oppression and abuse perceived by incels. This is mainly due to the nihilistic nature of the incel communities, which makes community members more likely to harm themselves than to take violent action on others to change their social environment. On the other hand, Zimmerman et al. (2018) argue that incels see themselves as “victims of oppressive feminism, an ideology which must be overthrown, often through violence”.

Bratich and Banet-Wiser (2019) argue that incels are, above all, the result of failure. Prevailing neoliberal ideas promote that achieving success requires mastering certain technical skills, such as picking up women. Incels fail to master these skills and to “entrepreneurialize themselves” to be able to attract women, which leads to failures in picking up women and, eventually, to the loss of confidence. As Baele et al. (2019) put it, incels have created different social categories for individuals (such as Chads and Stacys) that are seen constant and unchanging. In other words, incels believe that individuals cannot climb the social hierarchy ladder. This is why incels, as the lowest group in the hierarchy, are unable to form any romantic or sexual relationships with women. According to Bratich and Banet-Wiser (2019), neoliberalism itself cannot manage its failures since incels are unable to restore their confidence and wind up behaving hostile towards women.

Vito et al. (2017) have studied the relationship between the concepts of masculinity and violence. Their study focuses on analyzing Elliot Rodger’s online manifesto. Vito et al. argue that because of his characteristics (such as short height, muscle weakness), Rodger felt that he did not meet the standards of masculinity that were imposed on him by society. He also did not receive societal confirmation of his masculinity despite his efforts (e.g. spending time doing his hair). Rodger went through a crisis of masculinity and started to direct his feelings of anger toward those who he thought were lower on the social hierarchy, particularly women. He then adopted a violent and “true” masculinity to prove his manhood.

As stated above, the incel communities have praised Rodger’s actions, and he is still considered a hero in the incel online communities (Vito et al. 2017). Rodger can be seen as a part of the incels’ “lineup of ‘Saints’” that includes members of the community who have engaged in violent attacks for the good of the ideology (Baele et al. 2019). Vito et al. (2017) argue that the worship of Rodger in the online communities indicate that, just like Rodger, many incels feel pressure to uphold hegemonic masculinity standards. Maintaining hegemonic masculine ideals put us all at risk of violence, which should be recognized especially with regard to younger generations.

## 4. Analysis

In this part, we will discuss the question of the echo chambers fueling the incel movement in the digital era. How well suited is the theory to explain this phenomenon? What kind of criticism has the theory faced? Will the echo chamber theory help us understand better the emergence and the dynamics of the incel movement?

The echo chamber theory, as discussed in the earlier sections, includes the idea of online discussions taking place in closed “chambers”, where people surround themselves with others sharing the same thoughts and values as them. This theory has not been applied to the research of the incel culture before. According to Karlsen et. al. (2017) the echo chamber theory has been criticized for not being sufficiently able to explain the logic of online debating and behavior in general. According to this research, people tend to become more certain about their own opinions after Internet debates with those who disagree with them (Karlsen et al. 2017).

One could ask whether incels actually try to avoid different opinions or whether they seek out and then attack different opinions and the people presenting them. The idea of the trench warfare dynamics of online debates presumes that the opposite opinions and arguments actually fortifies individuals’ existing opinions. Also, if the opinion or belief of someone is already very intense, so is that person’s will to defend it. (Karlsen et al. 2017)

There are many examples of incels trying to actively silence unwanted people, e.g. women, by using aggressive messages and insulting language towards them (Jaki et al. 2019). There are also many occasions where incels have found women outside their own community and platforms and attacked them verbally. Is this just an outcome of being surrounded by similar thoughts, as in an echo chamber, or is it something more?

In addition, the idea of the spiral of silence seems to be inadequate to explain the incel movement. As mentioned in the previous section, the main argument is that people tend to stay silent rather than reveal their divergent opinions since they are afraid of becoming isolated from the rest of society. Yet the basis of the incel movement is the shared experience of not-belonging and already being in a way isolated from the world of Chads.

We can see a broader pattern of growing misogyny in the past years (Jaki et al. 2019). The movement can be seen as part of our popular culture, appearing in the language used by politicians, in justifications for changes in abortion legislation, as well as in terrorist attacks towards women. Feminist theory sees the incel culture as part of a larger rise of old-fashioned patriarchy (Higgins 2018). We could view these cultural patterns as not just a part of the incel culture, but actually a very fundamental feature of the movement.

The misogynist idea of women being especially the sexual property of men can be tracked back to the Victorian era (Collins 2018-2019). The idea that men have the access to the female body whenever they feel like it, is something very much underlying in the incel culture, too, and thus seems not to be something particular just for the digital era. In their essay, Brooke Collins brings up the prospect of violence against women committed by incels as not something new and unusual in our societies. Instead, Collins sees patriarchal violence towards women, who challenge their designated sexuality and sexual roles, as a phenomenon that has existed for centuries. A crucial part of the incel culture is the notion of something utterly wrong with the free choice of women and the free expression of female sexuality. Incels see men as inherently superior to women, and women existing only for the sexual pleasure of men (Collins 2018-2019). The new digital tools have of course helped spread these ideas, but can we say they have fueled them? Is the increased number of misogynist attacks inspired by others online or by normalizing the misogynist language everywhere else in society? This might be the crucial question in understanding the incel culture and in evaluating whether the echo chamber theory explains it: do misogynist ideas spread in echo chambers, or is society accepting this type of language more generally, at other levels, as well?

Some scholars have argued that the incel culture is part of the rising alt-right movement in the US and elsewhere. This is a complex phenomenon entangled with evangelical Christianity, corporate interests and media, e.g. the 4chan forum (Michelsen & De Orellana 2019). But there can be seen a correlation between the rise of the extreme right and the increasingly violent incel culture. It is important to notice that even though not all mass killings in, for example, North America, are committed by incels, Bratich and Banet-Weisen point out: “since 2007 in North America, many mass killings have been claimed by them [incels], and almost all are White” (Bratich & Banet-Weisen 2019).

It seems not adequate to look at the incel culture as a separate movement of lonely men isolated from the rest of society. The alt-right movement sees cultural liberalism as a hegemonic ideology which the members of the movement want to resist. The main focus is not only to form a group identity by sharing misogyny and other kind of hatred towards different groups of people, but the movement is about resilience and resistance (Michelsen & De Orellana 2019). The incel movement seems to be a part of this broader “critique” or “resilience” towards cultural liberalism, focusing primarily on its gender ideology. Incels accuse the modern gender ideology of disrupting human nature and their resistance is shown e.g. in the language and words they use, such as “feminazi” (ibid). The idea of resistance and resilience seems to be something more active than just staying in a chamber listening to one’s similar thoughts and views echoing from the walls.

Nonetheless both the Alt-Right and incel cultures use the same platforms – e.g. 4chan and Reddit – as well as similar language, memes and other shared cultural concepts (Daniels 2018). There have been far-right extremist terrorist attacks wherein the attacker has explicitly named the Internet as being an important element in their radicalization (Quek 2019). On the other hand, far-right ideology is generally based on perceiving a threat (ibid.). Can we say the same about the incel culture? At the end, the Internet has been an important element in spreading the extreme ideas of these two phenomena and it has enabled the attackers to share their thoughts and manifestos with a broader audience. In the most unfortunate occasions, this has inspired more mass murders. The importance of the different online platforms must not be underestimated when researching the incel culture. But are the ideas formed online and then spread elsewhere, or is the Internet just another location where growing misogyny and far-right ideology can be spread?

## 5. Policy recommendations

Based on our analysis there are a number of policies that could be applied in order to tackle the incel phenomenon. To clarify, the problem that these policies could fix refers to the social conditions where certain individuals feel such anger and resentment towards the surrounding society and its members that they would resort to extreme, violent measures, not the phenomenon itself per se. As our main argument was that echo chambers fuel the phenomenon, the solutions lie in the digital platforms’ handling of these chambers. In this case, reductionist strategies would be appropriate, and enhancing community rules and moderation on these digital platforms should be considered.

The problem is that even though these platforms, for example Reddit and 4chan, are monitored and moderated, this is quite difficult because even as a comment or a thread is moderated another one pops up. These platforms are usually based on unilateral moderation, which means that a few community members are chosen to act as moderators and go through the conversations. They can use "automoderation" as a tool to help them make their job more efficient. This basically means that moderators can apply filters, i.e. key words to find comments that do not follow community guidelines (Renfro 2016). These filters are easy to trick, however, by using euphemism or slang – something that is already common in the incel community – making them less efficient.

Some anonymous digital platforms, such as Jodel, have used user moderation. This means that a significant portion of community members are given moderation rights who then review reported posts. The moderators' decisions are based on the post at hand, not the user. A moderation algorithm then calculates how many moderators are needed to reach a decision and there is always a minimum number of moderators needed – no moderator can decide alone whether the post is banned or not (Jodel 2017). User moderation is an intriguing idea, but even though Jodel has had positive experiences with this system, it would not be a suitable solution to tackle the incel issue, as the main problem is that like-minded people gather in their own threads or platforms.

Automatic filters are also in use in some digital platforms, for example Facebook. However, automatic filters can be seen as too restricting, as these platforms rely on the content users create. Hence, automatic moderation is seen as a way to diminish users' freedom and democracy in the platform (de Zwart 2018). If community rules and standards are clearly stated, we would not regard automatic moderation as a problem. However, if we consider the incel phenomenon, this might not be the most efficient solution: we have seen in the past that if one platform gets too restrictive, users will find another platform (for example when the more radical incels moved from 4chan to 8chan, which then later got shut-down altogether).

When it comes to the reductionist approaches, we still consider the use of algorithms and machine learning as tools to moderate digital platforms more efficiently to offer the best solutions for the more extreme forms of the incel phenomenon. The popular anonymous platform Reddit has already implemented some machine learning tools to support their moderation, but these are merely tools that helps prioritize more urgent reports (Robertson 2019). Reddit also took action when it comes to enhancing community guidelines, as when they implemented a new policy and banned one popular incel thread (Hauser 2017). Together with clear community rules, the continuous evaluation of the adequacy of the rules, as well as the wider use of algorithms and machine learning, we believe that the most extreme forms of the incel phenomenon can be more easily detected and in the best-case scenario violent hate crimes can be prevented.

However, as our analysis stated, digital platforms do not fully explain the incel phenomenon even though they work as means to spread the incel message and, in worst cases, manifestos before mass killings. It is important to consider the moderation of these platforms but as surrounding society plays a significant role in creating an environment where misogyny and hate speech is tolerated, an even better way to answer the problem are more holistic policies considering society as a whole.

Zimmerman et al. (2018) suggest that misogynistic ideology ought to be addressed with the same seriousness as other forms of violent extremism. Violent attacks by incels have often been dismissed in the media as random acts of violence. Even at the government level – especially in the United States – attacks have been claimed to be the result of mental illness if the perpetrator was a white male. This discourse needs to change in order for the phenomenon to be taken seriously. Zimmerman et al. (2018) encourage implementing policies against hate speech and clearly sanctioning people who try to incite violence or harm against others with their speech. It is one thing to have your thread deleted from Reddit where you can anonymously write basically anything, but quite another to have a real fear of the authorities getting involved. However, it is hard to see these kinds of restrictive policies being implemented in the land of the free. If we consider the United States, a more appropriate policy recommendation would be stricter gun control. Zimmerman et al. (2018) suggest that one option could be closing background check loopholes that make it possible for individuals who are prohibited from buying guns to purchase them anyway.

Another way to approach the incel phenomenon is through education. Some aspects of the ideology might be addressed in schools by trying to curb misogynist ideas in the early stages rather than trying to block conversations on online platforms later on. This would demand changes in the curriculum and encouraging diverse dialogue in the classrooms. Education could also address the importance of healthy sexual culture and healthy ways of expressing one’s sexual needs. Healthy sexual culture could also be promoted at a societal level, by for example making sex toys and dolls more available and the use of them more accepted. This could be done through sexual education in schools as well as promotional campaigns lead by NGO’s or the government. The government could also address the issue of loneliness among young men by subsidizing different services that offer physical engagement. “Hug as a service” is already a popular concept in the United States (Tikkanen 2017) and the government could promote similar, non-sexual, low-threshold services.

## 6. Conclusion

In this paper, we have considered how digitalization has fueled misogynist movements over the last ten years. In the digital era, different misogynist movements have blended into the incel culture that is characterized by hostile behavior towards women and resistance to liberal values. It is evident that women have been subject to harassment and violence throughout history. However, technological development has created platforms where like-minded individuals can share their views and see themselves as communities. These online platforms often work as echo chambers where certain ideas are reinforced and opposing opinions are suppressed.

Following the development of different online platforms, misogynist communities have taken more organized and extreme forms.  Over the last few years, communities have shifted from words to action and many violent attacks have been committed by individuals who identify themselves as members of incel communities. In many cases, decision-makers, public and scholars have brought up the responsibility of online platforms in these violent actions. Some platforms have changed their policies towards stricter moderation (e.g. Reddit), whereas some platforms have been completely removed (e.g. 8chan).

We find that platforms' stricter moderation policies can diversify discussion to some extent, which in turn, could reduce the most extreme views and actions. However, we have also found that incels have been able to reorganize from one platform to another when moderation policies were changed. This also indicates that incels are particularly looking for echo chambers, where they can express their opinions freely. Therefore, echo chambers theories alone do not completely explain why the incel movement has grown so rapidly over the last ten years.

We discovered that the incel culture is closely related with broader movement that resists prevailing liberal culture. The so-called alt-right movement also has its roots in online communities from which it has found its way into public debate. We find that closing or strongly moderating online platforms, that work as echo chambers, will not tackle the issues that are the building blocks of these movements. Decision-makers need a deeper understanding of how surrounding society creates an environment wherein certain group of individuals feel anger towards other groups. A better understanding of these issues will guide us to find solutions through different policies and education.

## References

Alloway, T., Runac, R., Qureshi, M., & Kemp, G. (2014). Is Facebook linked to selfishness? Investigating the relationships among social media use, empathy, and narcissism. *Social Networking, 3*(03), 150-158.

Baele, S. J., Brace, L., & Coan, T. G. (2019). From “Incel” to “Saint”: Analyzing the violent worldview behind the 2018 Toronto attack. *Terrorism and Political Violence*, 1-25.

Beauchamp Z. (2019). *Our Incel problem*. Retrieved January 15, 2020, from https://www.vox.com/the-highlight/2019/4/16/18287446/incel-definition-reddit

Beauchamp Z. (2018). *Incel, the misogynist ideology that inspired the deadly Toronto attack, explained*. Retrieved January 15, 2020, from https://www.vox.com/world/2018/4/25/17277496/incel-toronto-attack-alek-minassian

Bratich, J. & Banet-Weiser, S. (2019). From Pick-Up Artists to Incels: Con(fidence) Games, Networked Misogyny, and the Failure of Neoliberalism. *International Journal of Communication, 13*, 5003-5027.

Collins, B. (2018). A Horror Tale of Male Entitlement: Jack the Ripper and ”His” Shadow, the Incel Movement. *Institute for Public Policy Research Journal, 13*, 10-16.

Collins, B & Zadrozny, B. (2018). *After Toronto attack, online misogynists praise suspect as 'new saint'*. Retrieved January 15, 2020, from https://www.nbcnews.com/news/us-news/after-toronto-attack-online-misogynists-praise-suspect-new-saint-n868821

Duke, A. (2014). *Timeline to 'Retribution': Isla Vista attacks planned over years*. Retrieved January 15, 2020, from https://edition.cnn.com/2014/05/26/justice/california-elliot-rodger-timeline/

Granovetter, M. (1978). Threshold models of collective behavior. *American journal of sociology, 83*(6), 1420-1443.

Hauser, C. (2017). *Reddit Bans 'Incel' Group for Inciting Violence Against Women*. Retrieved January 15, 2020, from https://www.nytimes.com/2017/11/09/technology/incels-reddit-banned.html

Higgins, C. (2018). *The age of patriarchy: How an unfashionable idea became a rallying cry for feminism today*. Retrieved January 15, 2020, from https://www.theguardian.com/news/2018/jun/22/the-age-of-patriarchy-how-an-unfashionable-idea-became-a-rallying-cry-for-feminism-today

Jaki, S., De Smedt, T., Gwóźdź, M., Panchal, R., Rossa, A., & De Pauw, G. (2019). Online hatred of women in the Incels.me forum: Linguistic analysis and automatic detection. *Journal of Language Aggression and Conflict*, *7*(2), 240-268.

Karlsen, R., Steen-Johnsen, K., Wollebæk, D., & Enjolras, B. (2017). Echo chamber and trench warfare dynamics in online debates. *European Journal of Communication*, *32*(3), 257-273.

Massanari, A. (2017). #Gamergate and The Fappening: How Reddit’s algorithm, governance, and culture support toxic technocultures. *New Media & Society*, *19*(3), 329-346.

Michelsen, N., & De Orellana, P. (2019). Discourses of resilience in the US alt-right. *Resilience*, 1-17.

Noelle-Neumann, E. (1986). *The Spiral of Silence: Public Opinion - Our Social Skin.* Chicago: The Univ. of Chicago Press.

Sageman, M. (2008). *Leaderless Jihad: Terror Networks in the Twenty-first Century.* Philadelphia: University of Pennsylvania Press cop.

Sunstein, C. R. (2009). *Going to Extremes: How Like Minds Unite and Divide.* Oxford, NY: Oxford University Press.

Sunstein C. R. (2017). *#Republic – Divided Democracy in the Age of Social Media.* Princeton and Oxford: Princeton University Press.

Van Alstyne, M. & Brynjolfsson, E. (1996). *Electronic Communities: Global Village or Cyberbalkans.* Cambridge: MIT Sloan School.

Vito, C., Admire, A., & Hughes, E. (2018). Masculinity, aggrieved entitlement, and violence: considering the Isla Vista mass shooting. *NORMA, 13*(2), 86-102.

Williams, Z. (2018). *‘Raw hatred’: why the 'incel' movement targets and terrorises women.* Retrieved January 15, 2020, from https://www.theguardian.com/world/2018/apr/25/raw-hatred-why-incel-movement-targets-terrorises-women

Zimmerman, S., Ryan, L. & Duriesmith, D. (2018). Recognizing the violent extremist
ideology of ‘Incels’. *Women In International Security Policy Brief*, 1-4.

## 5.2 #MeToo Movement, Digital Media and the Public Sphere

Eleanor Suovilla, Pietari Suomela, Anniina Riikonen, Susanna Kupiainen, Anni Juusola
Faculty of Social Sciences, University of Helsinki

## Abstract

In this paper we will examine the influence digital media has had on political dialogue in the public sphere. We will explore the phenomenon through an example case, namely the global feminist #MeToo movement which started in 2017. Within the framework of the #MeToo Movement, we introduce and examine the challenges digital media poses to the political dialogue in the public sphere. We start by going through concepts and theories utilized in this research paper. Then we will discuss the relationship between digital media and #MeToo, after which we will assess the negative and positive outcomes of the #MeToo movement. Finally, an overall assessment concerning the movement and phenomenon around it is given. Our main argument is that while making the public sphere more inclusive, digital media has also made public debate and political discussion more polarized and antagonistic.

*Keywords*: Public sphere, digital media, #MeToo movement, Twitter, social movements, rape, activism, collective action

This study attempts to describe how digital media has affected public debate. We will use the #MeToo campaign as a case example to show the impacts — both negative and positive — that digital media has had on the public sphere. The campaign can be described as a form of mobilization that initially was meant to draw attention to the scope of how much women still face sexual harassment and violence. The aim was to empower women to speak out, so that people would recognize the problem as relevant. Although the campaign succeeded in making women speak up and certainly drew attention to the matter worldwide, the reception was not completely positive. An opposing side emerged that questioned the campaign's endeavour for equality. We will look at the different types of outcomes in more detail in sections four and five.

As a theoretical background we will utilize studies that look into public discussion and the effects on it caused by digital media. We will reflect on the Habermasian public sphere, deliberative democracy and rational communication to see how well the Habermasian ideals of public debate are realized in the age of digital media. In addition, we will look into theories that criticize Habermas and show the difficulties of them actualizing, especially due to the emergence of social media.

## Theoretical background — Anniina Riikonen

It is useful to start our examination from Jürgen Habermas to uncover the meaning of the public sphere and deliberation. The ideal state for democracy for Habermas is deliberative democracy. Deliberation is an expression of people, together, contemplating issues that concern everyone, and deliberative democracy depends on this operation as its main principle (Bächtiger et al. 2018, 2). Voting then is not the only form of participation, as deliberation is a key factor affecting voting behaviour through rational communication (Bächtiger et al. 2018, 2). Many of the first theorists of deliberative democracy, most prominently Habermas, have also portrayed deliberation as being rational, free from power relations, aiming for common good, and being open and non-prejudiced towards different participants (Bächtiger et al. 2018, 3, 5). Thus, discussion participants do not blindly follow their own interests, but are willing to change their views. Later on, theorists assumed the debate situation as not being equal but pursuing inclusion of diverse groups (Bächtiger et al. 2018, 4).

A clear definition of ideal debate in deliberative democracy is communicative rationality. Communicative rationality is how debate in liberal democracies should be, and it is communication that forms through open-minded and reasoned discussion among equals (Cammaerts 2007, 3–4).

Deliberation takes place in the public sphere. The form of the public sphere changes in a historical process, wherein society and democracy are in transformation. Habermas describes the transformation of the public sphere during the 18th century as follows. As the economy changed along the emergence of capitalism, so did the social structure (Habermas 1999, 14). Finance shifted to a commercial private sphere external to individuals' households (Habermas 1999, 19–20). At the same time the press emerged, as a participant in commerce, to spread news to the public; and at the centre stood the bourgeois class (Habermas 1999, 21, 23). The state, as public authority, had the power to regulate markets. The interplay between the state and markets affected the public, and vice versa. This interplay gave space to public reasoning and critique (Habermas 1999, 24). Between private and public authority was the sphere in which the public sphere formed, containing the press, civil society organizations and public participation in politics (Habermas 1999, 30). Public opinion in the public sphere attempted to find just and right solutions through rational deliberation, and in the background the ideals of freedom and equality were forming (Habermas 1999, 54).

The Habermasian view of communication in the public sphere as being rational, open and aiming at consensus has since been subject to critique. Mouffe for instance (1999) has criticized this view of public debate for being consensus oriented. She points out that the public sphere has never been completely equal, nor free of conflicts — on the contrary, contradictions are part of democratic public debate (Mouffe 1999, 756). The polarization of the debate on sexual rights is also an example of the public not achieving consensus. Habermasian deliberative democracy has also been criticized for overestimating human capabilities for reasoned argumentation (Bächtiger et al. 2018, 20).

Not only have Habermas' original theories been subject to critique, but the very nature of communication has changed with the emergence of new technologies, such as the Internet. Especially social media, which was the platform the #metoo campaign, has changed participation. It can be claimed that participation is now open to a wider audience through the Internet. The new media have increased the participation of even those who did not participate before (Margetts et al. 2016, 157). However, making more people participate does not necessarily mean that communication or methods for communication are any more equal or more reasoned. Also, information is now more easily accessible. Thus, we need to explore more articles that reflect on the impacts of the new media on communication, deliberation and the public sphere. We will look into these theories later in this study, but to give some insight, it is useful to provide some background to the issue already at this point.

Public discussion on the Internet is not in accordance with Habermas' conception of public deliberation, even if it does produce new ways to participate (Dahlgren 2005, 151). As Dahlgren describes the communication patterns on the Internet: "The kinds of interaction taking place can only to a small degree be considered manifestations of the public sphere; democratic deliberation is completely overshadowed by consumerism, entertainment, nonpolitical networking and chat, and so forth" (2005, 151). According to Dahlgren, the benefits of the Internet as a new technology for the public sphere are in its effectiveness in including and forming a variety of interest groups and in that way developing multiple public spheres for public discussion (2005, 152). However, there is a problem with public spheres forming different groups: "…cyber ghettos threaten to undercut a shared public culture and the integrative societal function of the public sphere, and they may well even help foster intolerance where such communities have little contact with — or understanding of — one another" (Dahlgren 2005, 152). This habit of being too close to one's own group without discussing views with people from other groups is one possible explanation for why the #metoo campaign led to such a polarized debate.

We will come back to the reasons for this polarization later. With the new public spheres there might also be difficulties in trying to maintain a relationship between the multiple discussion spaces and institutions (Dahlgren 2005, 152–153). The impact of the Internet and social media on the public sphere has split theorists into two different camps. On the one hand, the impacts are seen as unimportant and not able to make a real difference to decision making. On the other hand, the impacts are seen as grand, changing the very nature of democracy by altering social structures and power relations in the global arena. (Dahlgren 2005, 154.)

## Digital media and the #MeToo movement — Pietari Suomela

This section examines the concept of *digital media* in the context of the #MeToo movement. One of the main points of this section is the notion that technological innovations are always neutral when first introduced. To understand the consequences which technological innovations have, they must be examined in a social context. There is a two-way road between technological innovation and social life, where both have an effect on one another.

### The concept of digital society

Simon Lindgren argues in his book *Digital Media & Society* (2017) that we can no longer make the distinction between the concepts of *digital media* and *digital society* (Lindgren 2017, 3). This statement, presumably, can be taken to refer to primarily post-industrial and relatively wealthy nations. Indeed, Lindgren presents a stack of comparable idioms for digital society: post-industrial society, information society and network society (Lindgren 2017, 4). Modern society is so saturated by “digital things” that it is getting increasingly more difficult to make the distinction between the terms *digital media* and *digital society*. However, the fusion of digital media and digital society is not self-evident (Lindgren 2017, 3).

Just like any groundbreaking technological innovation, technology associated with digital media has also influenced societies irrevocably. But it would be misleading to think that the relation between innovations and societies is a one-way road. Kranzberg’s first law, named after historian of technology Melvin Kranzberg, crystallizes this thought. First of all, technology is neutral. Secondly, technology interacts with society so that the consequences exceed the initial purpose of the technology. Thirdly, technology can have different results depending on the context and circumstances in which it is used. (Lindgren 2017, 4).

All three parts of Kranzberg’s first law are consistent with the concept of digital media. Especially the third part about different results of technology can easily be applied here. Digital media was an essential instrument in many of the largest events of the 2010s. The Arab spring started in 2010, Trump’s 2016 presidential campaign, and the

#MeToo movement started in 2017 were all very different kinds of events in relation to each other, but they were all made possible by digital media.

Lindgren also discusses the concept of media. This is an essential part of his argument, because digital media and digital society are redefining the concept of media. Furthermore, to grasp his argument about digital society fully, it is necessary to acknowledge the strong link between media and society. According to Lindgren, media is at the center of interaction between individuals and society and therefore it is quite easy to accept the fact that media plays an essential role in people's life and in the formation of an individual's self-portrait (Lindgren 2017, 5). Media structures, including languages and ways of thinking constrain and enable human interaction and should be studied if one tries to understand the so-called social reality (Lindgren 2017, 5-6).

Lindgren gives an illustrative example on how new communication technologies shape and define society: Just like the innovation of writing about 5000 years ago changed society dramatically and far beyond writing's initial purpose, so is digital media now (Lindgren 2017, 6-7). It is important to note that in both cases the road goes both ways: Communication technology shapes society, and individual behavior, just like the use of those technologies by individuals, shapes technology (Lindgren 2017, 7). Research that tries to explain how new communication technologies impact social lives and society as a whole are vitally important. A proper understanding of this is important especially in the early phases of new technologies, such as digital media.  This is because it is in those early stages that individuals and societies integrate new technologies into their everyday lives (Lindgren 2017, 7). If we lack an adequate understanding of the ways digital media has impact on our society, the consequences might be unpredictable and undesirable.

Social media can be understood as a subcategory of digital media. In essence it is a new kind of social dimension made possible by digital communication technologies. A logic called *networked individualism* is a good way to describe interaction through social media: Networks through which people interact are individually centered, looser, more open and more diverse than before, and digital media enables interaction between individuals in those networks. (Lindgren 2017, 27-28).

### Digital media and the #MeToo movement

The #MeToo movement, started in 2017, can be described as a social media campaign that successfully raised awareness and mobilized people in matters of gender, power and violence (Lindgren 2019, 2). What this section sets out to investigate is the role that social media had on making the movement possible. The aim is not to assess whether the movement or social media are good or bad, but rather to shed some light to the mechanisms behind social media movements in general.

First of all, it is good to remember that social movements are not new phenomena. What is new is that with the emergence of digital media and digital society, the blueprint of social movements has changed significantly. For example, the ways in which movements are actualized and people are being mobilized have radically changed along the emergence of a digital society. A general consensus among online social movement researchers is that the impacts of online movements are to be evaluated on case-by-case basis. The reason for this is that communication technology as such has no universally predictable consequences (Lindgren 2019, 3). The diversity of consequences is most easily understood when comparing two separate events in which social media campaigns played an essential role: the Arab spring in 2010 and #MeToo in 2017. The former resulted in regime changes in Arab countries, while the latter led to a widespread discussion about gender equality.

#MeToo is an example of hashtag activism which has been visible in political activism starting with the 2010s (Lindgren 2019, 4). Twitter as a real-time and global forum is the main platform of "hashtag-mediated public sphere", and hashtags themselves are tools to frame certain phenomena (Lindgren 2019, 4). For example, #MeToo is not just a reference to a phenomenon but also an indication of meaning and a term that frames the issue at hand. #MeToo has a semiotic function as defining a social phenomenon that would be hard to define or to name without using the term #MeToo.

In a way, digital media offers new tools to construct social life not just as a platform, but in more fundamental ways. Digital media influences directly how people act in public and even to some extent in private. Digital media and #MeToo question the distinction between private and public, and in this sense has a substantial influence on public discussion, individualism and privacy. The question of the relationship between private and public, a fundamental question in feminism, is probably one of the reasons why #MeToo has had a dividing effect.

Lindgren identifies three challenges for the #MeToo movement and social media movements in general: (1) noise and dilution, (2) hate speech and trolling and (3) clicktivism and disengagement (Lindgren 2019, 2). The #MeToo Twitter discourse became noisier and more off topic as it went on. This is not a surprising result as it is in line with both pre-digital and social medias' logic, in which focus on one particular topic is brief and quickly replaced by new topics (Lindgren 2019, 10).

As far as hate speech and trolling are concerned, conversation around the movement became more antagonistic, aggressive and negative (Lindgren 2019, 13). Political discourse is usually adversarial, but it seems that Twitter as a platform takes this antagonism further, making it difficult to have a sound political dialogue on the platform.

As time went on, there was a clear decrease in active participation in the #MeToo conversation on Twitter. However, participation activity in the #MeToo conversation still exceeded the activity of normal (non #MeToo or other campaign-like use) of Twitter use (Lindgren 2019, 15).

## Positive outcomes of the #MeToo movement — Susanna Kupiainen

The #MeToo-campaign is a very visible example of a phenomenon described as "hashtag-mediated public sphere" (Rambukkana 2015, 4). Political activism has taken to Twitter, likely because it is a global, real-time social media (Lindgren 2019, 4). The #MeToo followed the footsteps of #BeenRapedNeverReported (Mendes et al. 2018) and developed into a huge and controversial campaign that was widely covered by traditional media and noted at the highest levels of governance, with the Finnish parliament and president Trump, among others, voicing their opinions on the campaign. The purpose of this section is to examine the positive outcomes of the movement.

**Hashtag-mediated public sphere**

Combining hashtag-mediation and the public sphere for the concept of hashtag-mediated public sphere suggests that hashtag-oriented Twitter has become a new sphere for public discussion. The #MeToo campaign started by trending on Twitter, and soon spread into other social media as well. While Twitter is recognized as a hostile and aggressive environment particularly for feminist women, many feminists found participating in anti-sexual violence campaigns easier online than in their day-to-day lives (Mendes et al. 2018, 243–244). The ability of Twitter to support discussion and activism of these sensitive issues is a positive development.

Offering the possibility for anonymity behind a username, Twitter and other social media channels seem to be becoming increasingly important scenes of the global public sphere in western societies. Participating in these campaigns was not easy, but thousands of women and victims of sexual violence were given a voice and activism was celebrated by traditional media, bringing it to the attention of a much wider audience than only those on Twitter (Mendes et al. 2018, 244). The voice of women was heard loud and clear, considering that the topic was soon discussed even in the Finnish parliament across the Atlantic, with the Minister of Justice commenting that Finnish law condemns all sexual harassment, but people's attitudes and actions may not (Konttinen 2017). It was also suggested that the legal definition of rape should be changed to lack of consent, rather than defining it along the use of violence (ibid.).

Twitter as a public sphere for deliberative democracy seems to have done its job in this regard. Traditional news media understood that something was happening that many citizens wanted to change and started reporting it to those who do not use Twitter. Media coverage helped decision makers estimate the importance of the issue to citizens, fulfilling the democratic ideal of listening to the voices of the oppressed as a basis for decision-making.

Media has immense power with regard to setting the agenda and consequently determining what the public deems important. It can therefore shape preferences and opinions, and influence what people consider worthy of public discussion, or which social problems need to be solved (Flew 2018, 11–12). This was a huge part of the positive consequences of the movement, as they took place offline. Social media was the starting place for activism, but the societal change preceded online communities. The campaign was framed by the media mostly, especially in the beginning, in a positive way that supported the victims and raised concerns about the amount of unreported sexual violence in western societies where women are often thought to be quite safe.

### The culture of silence surrounding rape

Speaking out about sexual abuse "exposes the pattern of abuse, warns those who might become victims, and encourages others similarly situated to come forward with their own claims" (Prasad 2018, 2509). There has long been a culture of silence surrounding sexual abuse, which is made possible by the shame the victims feel for what happened and can be made worse by officials such as police suggesting that the victim was at fault too, since they should not have been drunk or dressed as they were. Many victims are afraid to report the abuse, as the campaign #BeenRapedNeverReported proved. The spreading of the campaign encouraged one person after another to be brave enough to make their painful experience public, which encouraged more people in return. The amount of people having suffered from sexual violence that were ready to go public with it was the basis for the campaign's powerful effects.

After a sexual assault, it can be easy to buy the silence of the victim with a non-disclosure agreement (NDA), especially in the United States. The victim feels alone, humiliated and scared that someone will find out, so they may be inclined to sign the immoral NDA, thus having to stay silent forever. The #MeToo campaign, and the women who spoke out about abuse despite having signed NDAs, have broken this silence induced by shame and fear. Speaking out instead of remaining silent was found so necessary that many states in the US started preparing bills that would limit the use of NDAs in sexual violence cases to let the victims speak out about them. It was also considered something the public should know about, to avoid being able to repeat the abuse in silence (Prasad 2018). Since the campaign, sexual harassment has been discussed more often and more openly, with many employers changing their harassment policies. For example, the congress of the United States added training, updated their complaint and counseling practices and increased the rights of unpaid interns (Prasad 2018, 2522–2523).

### Effects on rape culture

The #MeToo-movement was utilized not only to encourage women to speak up about sexual abuse, but also to attempt to make a change in toxic masculinity with regard to sexual violence, referred to as 'rape culture'. Rape culture is an attitude surrounding sexual abuse, characterized by silently accepting, excusing or even supporting acts of sexual assault (Pettyjohn et al. 2019, 1–2).

After the #MeToo-movement, several male-dominated hashtag campaigns were also started, with the hashtags #ItWasMe, #IHave and #HowIWillChange (Lindgren 2019, 3–4). The campaigns were a consequence of a shift in philosophical perspective on sexual violence, claiming men's responsibility in prevention of sexual violence (Pettyjohn et al. 2019, 2–3). While the backlash of the campaigns was very harsh and many found them ridiculous, there were still thousands of men genuinely reflecting on their toxic behavior, promising to be better in the future and most importantly, discussing how they can teach their children to be better (Pettyjohn et al. 2019, 3–8).

The narrative around the responsibility of those who may have not harassed anyone, but have silently accepted harassment, is a rather new one. “Locker room talk”, a term used by Donald Trump to justify offensive comments, has also been connected to toxic masculinity and rape culture by objectifying women and normalizing harassment-related speech. The narrative is a means towards a culture where offensive talk is not brushed off as ‘boys will be boys’, and men also have to take responsibility for their possibly innocent yet harmful words. Language is a consequence of attitudes and reshaping the next generation’s way of thinking begins with changing attitudes. These men-oriented hashtags show a valuable change in the attitude towards harassment and are valuable for the movement as such.

## Challenges and negative outcomes of #Metoo movement — Anni Juusola

This section addresses the challenges and negative outcomes of digital feminist activism and the #MeToo movement in particular. As the goal is to understand how digital media has affected the public sphere, the main focus is on the complex and problematic nature of the digital environment and the experiences of those who act within it.

Firstly, the impact of digital media on collective behavior is discussed from a critical perspective. The question of how the Habermasian public sphere and deliberative democracy are challenged by the digital revolution is tied to research on social movement. Finally, online abuse and the negative experiences of women who engage in digital feminist activism are examined.

### A critical perspective on social movements in the digital age

Research on social movements tries to answer the question of why social movements, such as the #MeToo movement, succeed or fail. Usually it is difficult to find a direct causal relationship between attempts of collective behavior and the final outcome (Carty 2015, p. 28). Indisputably, the #MeToo Movement gained substantial attention from mainstream media but researchers still know little whether or how this kind of hashtags can actually produce social change (Mendes, Ringrose and Keller 2018, p. 237).

In spite of these difficulties in analysing the outcome of the #MeToo Movement, it is clear that digital media offers new possibilities for all social movements. These days, activists can use social media platforms to raise awareness and organise events. They can reach large amounts of people quickly and challenge predominant views on an issue with their message.

Still some scholars are skeptical about these new possibilities created by digital media. It is claimed that people tend to interact with like-minded people online which can lead to fragmentation and polarization (Bimber and Davis 2013, p. 245). Carty (2015) refers to these phenomena as “cyber-balkanization” and the “echo-chamber” effect (p. 30). In other words, digital media and the Internet create small groups whose members share similar interests and despise outsiders with different views. These claims are contrary to the idealistic notion that digital media could potentially create “virtual public spheres" where people develop a sense of community regardless of physical distance (Kahn and Kellner 2003, p. 14).

The #MeToo Movement has been criticized by some as a "battle of sexes" which pits men against women (Kunst et al. 2019, p. 1). Unfortunately, there is still little knowledge why some groups perceive specific social media campaigns significant, while others find them harmful (Kunst et al. 2019, p. 6–7). Based on the findings of cross-cultural study on the underlying factors affecting men's and women's attitudes towards the #MeToo Movement, one way to reduce the polarization might be to highlight that campaigns such as the #MeToo Movement, raise awareness about sexual violence experienced by both men and women (Kunst et al. 2019, p. 20). Considering the framing of the campaign carefully might help to avoid the negative counter-reactions towards feminist digital activism.

Within a broader theoretical framework, these concerns give us some insight as to why digital media poses a threat to the Habermasian public sphere and deliberative democracy. Habermas has argued before the emergence of the new ITC that the mainstream media has had a negative impact on the public sphere. According to him, public opinion, which was once based on the outcome of debate and reflection, is now constrained by media experts who construct the public discourse to those themes they approve of (Carty 2015, p. 31). It could be argued that the rise of digital media and its negative side effects, like the "echo chambers" of the Internet, continue this trend.

Habermas also uses the concept of "ideal speech situation" in which communication is not controlled by political or economic forces and everyone participates in public debate on equal terms. Applying the ideas of Habermas, skeptical theorists think that virtual relations in cyberspace do not fulfil the conditions of the ideal speech situation. (Carty 2015, p. 31–32). For example, everyone does not have digital skills or access to technology to participate and the owners of the digital platforms also have their own economic and political interests which might prevent a truly equal public debate.

### Women's experiences of engaging in digital feminist activism

The rise of digital technologies has also enabled online abuse against girls, women and some men who participate in digital feminist activism. According to Citron (2014) some of the Internet's key features, namely anonymity, mobilization of groups and group polarization, make it more likely that people will act destructively. At the same time certain features, such as Google bombs, enhance the destruction's accessibility, making it more likely to inflict harm (p. 57). As interacting online can lead to fragmentation and polarization of opinions, it is no surprise that expressing feminist views may trigger vulgar counter-reactions.

Since 2014, Mendes, Ringrose and Keller (2018) have studied the experiences of organizers of feminist campaigns and those who have contributed to them by using hashtags, such as #MeToo and #BeenRapedNeverReported. Their approach to studying digital feminist activism is rather unique because they combine the perspective of how digital tools are used and the experiences of the users.

Mendes, Ringrose and Keller (2018) have focused on Twitter as a platform which also happens to be one of the main digital tools of the #MeToo Movement. Their findings indicate that negativity, hostility or trolling in response to expressing feminist views online is a common experience. Within their study of 46 active Twitter users who self-defined as “feminist activists”, 72% of the respondents had experienced online abuse. These experiences included a wide range of practices starting from mean comments, such as “you are ugly”, to multiple attacks on the activist’s Twitter feed or graphic rape and death threats. Notwithstanding the online abuse, most participants persisted in their digital feminist activism and developed strategies to cope with harassment. (Mendes et al. 2018, p. 242–243).

It is important to note that engaging in digital feminist activism can create mixed feelings among participants even though they would not encounter online abuse. The #BeenRapedNeverReported hashtag trended in 2014 and it was in many ways similar to the #MeToo hashtag. It was used by girls and women to share stories of sexual violence and why they did not report the assaults to authorities at the time. (Mendes et al. 2018, p. 237).

After analyzing hundreds of tweets with the hashtag #BeenRapedNeverReported and interviewing girls and women who had used it, Mendes, Ringrose and Keller (2018) found that participating in the #BeenRapedNeverReported hashtag was both a comforting and triggering experience. Many participants described how the hashtag evoked difficult and upsetting emotions although they also emphasized the importance of the support of other women and girls. (Mendes et al. 2018, p. 238). It is very likely that the user experience of the #MeToo hashtag would be very similar. Thus, it can be concluded that digital feminist activism has a complex nature and it is often challenged by misogynist views. Women, girls and men who engage in digital feminist activism are at risk of online abuse. Sometimes the activism itself might evoke consuming and difficult emotions.

## Overall achievements of the movement in terms of social capital formation — Eleanor Suovilla

This part of the essay will discuss the overall achievements of the movement in terms of social capital formation. The focus is on trying to reflect whether digital networks can affect social capital formation offline.

### Collective and connective logic of actions

According to Bennet and Segerberg (2012), when communication becomes a prominent part of the organizational structure there are two underlying logics of action: collective and the connective. The collective logic of action emphasizes how it makes no sense for a rational individual to contribute towards resolving a common problem if the final result is unclear or if there is an opportunity for free riding. The logic also requires more efforts in achieving a collective identification which in turn demands resources and a more extensive formal organizational structure. Out of the two logics this is the traditional one which is challenged by the logic of connective action.

The connective logic of action according to Bennet and Segerberg (2012, p. 11) "applies increasingly to life in late modern societies in which formal organizations are losing their grip on individuals, and group ties are being replaced by large-scale, fluid social networks." The core of this logic is *digitally networked action* (DNA) which highlights the significant role that personalized action has in post-industrial democracies. People want more direct opportunities of engaging and self-expression while simultaneously detaching themselves from formal organizations, ideologies or political parties. Grossi brings forward his definition of a democracy of the individualized citizens which is characterized as a "intertwining and permanent conflict among social systems and worlds-in-life, between government and cultures of civil society, institutional power and individual empowerment" (2015, p. 28-29). An interesting question is whether networks that are built according to a connective logic could still enhance the level of social capital offline even when the logic itself does not require the construction of a unified "we" online (Bennet & Segerberg, 2012).

Bennet and Segerberg (2012) underline that the connective logic is about personal expression achieved by sharing. The formed connective networks place technology at the core of their function as they see digital media as their organizing agent. The individualized citizen of the 21st century according to Grossi therefore utilizes these technologies as the basis of citizenship, the argumentative-deliberative discursive located online (2015, p. 28). Therefore, democracy is no longer about searching for consensus but rather about contention and self-empowerment. In a sense the entire online network that the #MeToo campaign has produced could be analyzed by placing the communicative processes at the center of attention as the #MeToo movement became globally known once the #MeToo campaign went viral on Twitter.

According to Blaschke, Schoeneborn and Seidl (2012) there is an alternative method of trying to understand what organizations are, what the role of communication within them is and how they construct meaning. They introduce the approach of *communication constitutes organizations* (CCO) in comparison to network analysis which puts individuals at the center of attention. The CCO is best suited to elucidate the meso or translocal level of organizations. By using this approach one can study how an organization emerges on the local level and becomes a larger entity on the translocal level by examining various communication episodes. Essentially the approach highlights communication as the constitutive part of an organization as it has the ability to bring forward the processual, historically situated and politically contested character of organizing (Blaschke, Schoeneborn & Seidl 2012). The notions above could be combined with the thoughts of O'Hallarn (2016) when thinking about the link between social capital generation, Internet technologies and communication processes as the building blocks of a network. O'Hallarn (2016) mentions that one way of thinking about the construction of social capital is to see it as a result of the digital, connected network itself. Gibson, Howard and Ward add that social capital can be measured "either by studying the aggregate levels of association in a population, or by fully enumerating the density and reach of a particular individual's network of associations" (2000, p. 5).

**Social capital formation**

When it comes to the #MeToo movement it becomes challenging to understand if the communicative processes in the digital environment have managed to increase social capital offline. Sajuria, vanHeerde-Hudson, Hudson, Dasandi and Theocharis (2015) study in their research whether social media has led to the formation of bridging and bonding capital. They present in their article Putnam's objections for such a process as he saw that social capital cannot be fostered in a digital environment. The authors conversely claim that social media could serve as a platform which would lower initial limits to communication such as gender, race or disability. Gibson, Howard and Ward (2000) echo accordingly that an increasing level of women have moved on-line in the UK and USA since 1998. Sajuria et.al respectively argue "that Twitter and Facebook discussions create social networks, operating under norms of trust and reciprocity, that are able to mobilize resources and information" (2015, p. 712).

Sajuria et al. conducted research on the "online social architecture of networks of Twitter connections and conversations" in order to find evidence for patterns of bridging and bonding social capital (2015, p. 735). They found that ICTs have the potential of forming bonding capital but bridging capital formation did not seem to form organically. They did point out that there is an element of intentionality that is required in bridging the social capital of online environments. People from within the networks need to engage as brokers in order to produce bridging ties between networks. They did highlight that further research is necessary in order to understand whether the content of those networks and connections can provide evidence of social capital formation offline (Sajuria et al., 2015). The question of whether online connectedness has effects on social capital formation and political activity in the real world is a very complicated one which cannot be answered in this part of the essay.

The #MeToo campaign certainly raised the public consciousness regarding sexual violence as the collective communication flows happening online had a spillover effect bringing the topic into national arenas of discussion and ultimately taking the discussion to a global level. There have been several positive outcomes worldwide as mentioned in the previous section. The individuals of the #MeToo campaign network therefore were the brokers of bridging social capital but the question that remains open is whether the content of their communicative processes furthered the bridging of social capital. In other words, according to Sajuria et. al (2015) a distinction has to be made "between the thinner, transactional view of connective action and the thicker, transformational view of social capital". This essay does not have an answer to whether the #MeToo campaigns communication network succeeded in bridging social capital offline. That being said, the campaign certainly serves as an interesting object of study in terms of conceptualizing how the connections may have the potential of forming positive externalities in the form of social capital (Sajuria et.al, 2015).

## Conclusions

Section two examined the relation between technological innovations and society through the example of digital media and the #MeToo movement. Just like with every technological innovation, digital media should also be examined as a part of society, not as a separate phenomenon. Digital media allowed the #MeToo movement to create and define a new social and political discussion. The fact that the #MeToo movement was created on social media is an essential part of the overall effects of the movement. The medium through which a discussion in society takes place has an effect to the final outcome.

As social activism, the #MeToo movement developed into a major campaign and discussion beyond Twitter and the Internet itself. Traditional news media helped spread the discussion across western societies. The movement had multiple positive consequences, as discussed in section three, mostly with regard to the culture of silence and rape. People speaking up about abuse encouraged other people to speak up, leading to a cycle of breaking the silence surrounding rape. The issue was taken seriously; several laws were proposed to be changed (in the U.S. alone), while many employers checked and updated their procedures on handling and reporting sexual harassment. The effects on rape culture were based on confronting men not only as abusers but also as silent bystanders. Many men also realized their own harmful ways and appeared dedicated to teach their children about the concept of consent.

Today, all social movements are faced with new challenges created by digital media and technology. Thus, the #MeToo movement also had its negative outcomes. Section four addressed the whole social movement critically. Within a broader framework, the question of how the Habermasian public sphere and deliberative democracy had been challenged by the digital revolution was also contemplated.

It could be concluded that certain features of the Internet enhance online abuse. Online interaction may also lead to fragmentation or polarization of opinions which might explain why digital feminist activism often encounters vulgar counter-reactions. Considering the framing of social movements, such as the #MeToo Movement is crucial in order to reduce the polarization of opinions and online abuse towards participants.

The fifth section reflected on whether the content of the communicative processes of the #MeToo campaign could create bridging capital in the real word. Even though an answer to the question is beyond the scope of the essay, it presents an intriguing topic of research. According to O'Hallarn (2016) the first step would be to identify if the public sphere can be proved to exist in digital environments in the first place. Further research could take place in order to identify if social capital offline could be created by the online public sphere. Next, it would be beneficial to determine if political activity directly results from the digital public sphere or if political activity has alternatively required the formation of the by-product of an online public sphere operating under the logic of connectedness, namely social capital.

## References

Bennet, L. & Segerberg, A. (2012). The logic of connective action: Digital Media and the Personalization of Contentious Politics. *Information, Communication & Society*, *15*(5), 739-768. http://dx.doi.org/10.1080/1369118X.2012.670661

Bimber, B. & Davis, R. (2013). Campaigning Online: The Internet in U. S. Elections. New York: Oxford University Press.

Bächtiger, André, John S. Dryzek, Jane J. Mansbridge, and Mark Warren, eds. (2018). The Oxford Handbook of Deliberative Democracy. First edition. Oxford, United Kingdom; New York: Oxford University Press.

Cammaerts, B. (2007). 'Citizenship, the Public Sphere and Media'. In Reclaiming the Media: Communication Rights and Democratic Media Roles, European Communication Research and Education Association series, Bristol: intellect, 1–8. Retrieved from: http://eprints.lse.ac.uk/39663/1/Cammaerts_citizenship_public_sphere_2012.pdf.

Carty, V. (2015). Social Movements and New Technology. New York: Routledge.

Citron, D. K. (2014). Hate Crimes in Cyberspace. Cambridge, Massachusetts: Harvard University Press.

Dahlgren, Peter. (2005). 'The Internet, Public Spheres, and Political Communication: Dispersion and Deliberation'. Political Communication 22(2): 147–62.

Flew, T. (2018). Understanding Global Media. 2nd edition. London: Red Globe Press.

Gibson, R., Howard, P., and Ward, S. (2000). Social Capital, Internet Connectedness & Political Participation: A Four-Country Study. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.11.8677&rep=rep1&type=pdf

Habermas, Jürgen. (1999). The Structural Transformation of the Public Sphere: An Inquiry into a Category of Bourgeois Society. 10. print. Cambridge, Mass: MIT Press.

Kahn, R. & Kellner, D. (2003). Internet Subcultures and Oppositional Politics. In The Post-Subcultures Reader, edited by Muggelton, D. & Weinzierl, R., p. 299–314. London: Berg.

Konttinen, M. (2017) Eduskunta kävi vakavan keskustelun seksuaalisesta häirinnästä – "Laaja rakenteellinen ja vastenmielinen ongelma" [news article]. Yle 12.12.2017). Retrieved from: https://yle.fi/uutiset/3-9973153

Kunst, J. R., Bailey, A., Prendergast, C. & Gundersen, A. (2018). Sexism, rape myths and feminist identification explain gender differences in attitudes toward the#metoo social media campaign in two countries. Media Psychology, 22(5), p. 818–843. Retrieved from https://doi.org/10.1080/15213269.2018.1532300.

Lindgren, S. (2017). Digital media & society. Thousand Oaks, CA: SAGE Publications.

Lindgren, S. (2019). Movement Mobilization in the Age of Hashtag Activism: Examining the Challenge of Noise, Hate, and Disengagement in the #MeToo Campaign. Policy & Internet. doi:10.1002/poi3.212

Margetts, H., Peter, J., Hale, S.A & Yasseri, T. (2016). Political Turbulence: How Social Media Shape Collective Action. Princeton, New Jersey: Princeton University Press.

Mendes, K., Ringrose, J. & Keller, J. (2018). #MeToo and the promise and pitfalls of challenging rape culture through digital feminist activism. European Journal of Women's Studies. 25(2), p. 236-246. doi:10.1177/1350506818765318

Mouffe, Chantal. (1999). ‘Deliberative Democracy or Agonistic Pluralism?’ Social Research 66(3): 745–58.

O’Hallarn, B. (2016). The public sphere and social capital: Unlikely allies in social media interactions? *First Monday, 21*(10). doi: http://dx.doi.org/10.5210/fm.v21i10.6961

Pettyjohn, M., Muzzey, F., Maas, M. & McCauley, H. (2019). #HowIWillChange: Engaging Men and Boys in the #MeToo Movement. Psychology of Men & Masculinity. 20(4), 612-622.

Prasad, V. (2018). If anyone is listening, #metoo: Breaking the culture of silence around sexual abuse through regulating non-disclosure agreements and secret settlements. Boston College Law Review, 59(7), 2507-2550.

Rambukkana, N., ed. (2015). Hashtag Publics: The Power and Politics of Discursive Networks. New York: Peter Lang

Sajuria, J., vanHeerde-Hudson, J., Hudson, D., Dasandi, N., & Theocharis, Y. (2015). Tweeting Alone? An Analysis of Bridging and Bonding Social Capital in Online Networks. *American Politics Research, 43*(4), 708 –738. DOI: 10.1177/1532673X14557942

# Part VI

# Democracy, Communication, Algorithmic Governance

14:01
23° Mostly Cloudly
FAKE NEWS
LIVE
4 ▲ +0,36% United Kingdom Pound 0,7242 ▼ -0,43% Gold 1,628.25 ▼ -0,056% Natural Gas 4.025 ▲ +0,24% Oi

# 6.1 Ideals and Agency in the Fight Against Misinformation on Online Platforms

Nico Stockmann, Judith Knebler, Hannamari Hoikkala, William E. Burden,
Anna M. Bogdan
Faculty of Social Sciences, University of Helsinki

## Abstract

While digital platforms remove barriers to accessibility and production, they nonetheless create new challenges that cannot be solved from a simple transference of traditional media regulations online. One of the most visible of these challenges is the spread of misinformation, which is considered a threat to democracy. This research paper examines how two government initiatives — Germany's *Network Enforcement Act* and the *Nordic Model* — tackle misinformation on social media and their impact on individual and collective agency in deliberative democracy. Paul Grice's *Cooperative Principle* and Jürgen Habermas' concepts of the *public sphere* and the *lifeworld* offer theoretical frameworks for exploring why truth-telling is vital to communication and democracy. We argue for a paradigm shift in the usage and expectations towards algorithmic-based regulation, and the strengthening of individual and collective agency through government initiatives that emphasize individual media literacy and encourage trust in traditional media actors.

*Keywords*: Misinformation, truth, communication, democracy, social media, online platforms, algorithm, Habermas, Grice, cooperative principle, public sphere, Nordic Model

While digital platforms remove barriers to accessibility and production, they nonetheless create new challenges that cannot be solved from a simple transference of traditional media regulations online. Perhaps the most visible of these challenges is misinformation. According to Shao et al. (2018, p. 2), the “massive spread of digital misinformation has been identified as a major global risk and alleged to influence elections and threaten democracies”. Though misinformation in the media or politics may not be a novel problem, “the ease with which social media can be manipulated” and the immense exposure achievable online are (Shao et al., 2018, p. 2).

Not only do alternative media sources face lower costs to produce misinformation, the existence of algorithms and bots — artificial intelligence utilizing machine learning to fulfill assigned tasks (Boulanin, 2019, p. 16) — present new threats as they can limit information exposure, promote false news, and lead to overzealous censorship (Shao et al., 2018). To grasp this challenge, this paper is based upon the theory of truth-telling proposed in Paul Grice’s (1975) Cooperative Principle as it pertains to Jürgen Habermas’ (1991) concepts of the public sphere and lifeworld. The former concept refers to citizens’ engagement in public debate as a means of participating in society and politics, and the latter refers to the intersubjective web of language and culture constantly operating at the background of culture (Habermas, 1987). We also consider S. M. Amadae’s (2018a) application of these theories to the digital revolution.

Through these lenses, we argue for a paradigm shift in the usage and expectations towards algorithmic-based regulation, and therefore a strengthening of individual and collective agency through government initiatives that emphasize individual media literacy and encourage trust in traditional media actors. The paper proceeds as follows: First, we highlight the necessity of truth-telling and intelligibility in communication through Grice’s Cooperative Principle and Amadae’s work on truth-telling norms. Second, we examine the question of misinformation through Habermas’ discursive theory of democracy and concept of the lifeworld, as well as the impact on individual agency as communication moves increasingly to online environments. Following that, a case study on Germany’s Network Enforcement Act will provide a practical example of the challenges associated with regulating online communication. The fourth section will first examine the issues regarding algorithm-based regulation, and the next section builds on this to discuss smart regulations based on the Nordic Model.

## Principles

### Communication and truth-telling norms: Grice’s Cooperative Principle in the public sphere

As media increasingly move to online platforms, which can result in large-scale digital misinformation, it is vital to preserve a foundation of truth-telling in communication. Though it may seem obvious that one should tell the truth, as argued by Paul Grice (1975), it is important to understand “the standard type of conversational practice not merely as something that all or most do IN FACT follow but as something that it is reasonable for us to follow, that we SHOULD NOT abandon" (Grice, 1975, p. 48; qtd. in Amadae, 2018a, p. 20). As S.M. Amadae describes, losing this foundation can lead to either "a breakdown of communication in polarized situations of fundamental conflict"; or leads to a normalization of strategic rationality — by which self-interests are pursued above all — "[which] accepts that every utterance must contain accurate, false, or ambiguous information dependent on the reward structure of interactions" (2018a, p. 3). This section seeks to explain why truth-telling is vital to communication using Grice's *Cooperative Principle* (updated by Amadae for the digital era), and how this, in turn, can inform issues of misinformation and the lack of algorithmic intelligibility.

Grice first distinguishes between formal and natural language. Formal language represents scientific inquiry wherein "there are objective conditions that must be satisfied for propositions to be true" (Amadae, 2018a, p. 18). In contrast, natural language contains elements that "cannot be precisely/clearly defined" and therefore cannot be concluded to be objectively true, and are thus up for interpretation (Grice, 1975, p. 42). For example, *conversational implicatures* are facets of natural language that "a hearer can work out from the *way* something was said rather than *wha*t was said" (Grandy & Warner, 2017). Unlike formal language, these conversational implicatures depend on context and interpretation because "what is said may be true, but what is implicated may be false" (Grice, 1975, p. 58). Yet, despite this lack of objective clarity, people are able to understand each other and communicate. According to Grice, this is because of the Cooperative Principle (CP), which assigns logic to natural language and emphasizes cooperation and truthfulness as vital.

Grice's Cooperative Principle states, "Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose of direction of the talk exchange in which you are engaged" (Grice, 1975, p. 45). Conversations are assumed to be "cooperative enterprises" (Grandy & Warner, 2017). By entering into a conversation, one naturally takes on a cooperative relationship in order to have a rational exchange of information. Grice's theory emphasizes "[maximizing] efficient, rational, and cooperative" communication (Braaten, 1991, p. 60). If conversations developed in disjointed/uncooperative ways, they would not be able to exchange information and would thus be irrational (Grice, 1975, p. 45; Amadae, 2018a, p. 17). To follow the CP, then, is to act reasonably in communication.

Another way of understanding the fundamental nature of cooperation is by examining Grice's understanding of 'meaning'. By meaning something, the speaker intends for the listener to respond in a manner that recognizes the speaker's intentions (Amadae, 2018a, p. 18; Grandy &Warner, 2017). 'Meaning something' is cooperative because it relies on recognition of intentions and the creation of a subsequent response. However, as demonstrated in conversational implicatures, 'meaning something' can form from unstated implications as opposed to something explicitly said. Because of this distinction between what is said and what is implied, other principles are necessary for these phenomena to occur without detriment to communication.

The following maxims exist under the CP: Quantity, Quality, Relation, and Manner (see Grice, 1975). These maxims are all linked through the "supreme Conversational Principle" of cooperation, as described by the CP (Grice, 1989, p. 370). It is thus first necessary to assume that the goals of conversation are cooperative (Grice, 1975, p. 49). Following this is the most relevant maxim: Quality. It states that the speaker must attempt to be truthful. The submaxims of Quality are:

i. Do not say what you believe to be false
ii. Do not say that for which you lack adequate evidence (Grice, 1975, p. 46).

All other maxims can only occur after the conditions of truthfulness under Quality are fulfilled (Grice, 1975, p. 46). It is only possible to exchange information or to cooperate if the participants are assumed to be telling the truth. This is not to say that people cannot lie, only that doing so deviates from the goals of communication. Or, as Habermas argues, one can only lie if the point of communication is assumed to be the transference of meaning (Amadae, 2018a, p. 20).

Though the mechanics of their theories of communication may differ, both Grice and Habermas emphasize the importance of truth-telling and intelligibility (Braaten, 1991, p. 60), with Habermas arguing that a cooperative understanding of communication is necessary for the lifeworld (Habermas, 1984, p. 337). These assumptions are vital in understanding why misinformation — especially on the scale allowed by the Internet — is harmful to the public sphere as conceived by Habermas (1991). If the public sphere is understood as formed through communication based on the assumptions of cooperation and truth-telling, then misinformation — particularly on the Internet's massive scale — is especially problematic given its erosion of the fundamental conditions underlying cooperation (Amadae, 2018a, p. 21). This erosion can be seen by treating communication exhaustively as a means to maximize self-interests (i.e. game theory's strategic rationality). If perpetuated on the Internet for political and economic aims, a strategic view of communication potentially serves to normalize lying (Amadae, 2018a, p. 22). There can be no assumption that actors speak truthfully, thus leading to the aforementioned breakdowns in communication. In contrast, a theory of communication based on cooperation and truth-telling as argued by Grice and Habermas is important because such a "commitment to truthfulness […] can help sustain a public sphere basic to civil politics" (Amadae, 2018a, p. 21).

The Cooperative Principle does not just inform how commitments to truthfulness are vital as society moves toward digital platforms, it is also applicable to artificial intelligence (Grandy & Warner, 2017) and algorithmic regulation of misinformation because of the importance of intelligibility in natural language. One can only maintain conversational implicature if at least the CP is being observed and if the speaker believes that the listener is capable of intuitively deducing what is being implied (Grice, 1975, p. 50). One must therefore understand all of the following: "the conventional meaning of the words used together with the identity of any references"; "the CP and its maxims"; "the context, linguistic or otherwise, of the utterance"; "background knowledge"; and that all aforementioned requirements are "available to both participants and both participants know or assume this to be the case" (Grice, 1975, p. 50).

One relevant example of such conversational implicatures in the context of algorithms is that of irony. In irony, "though some maxim is violated at the level of what is said, the hearer is entitled to assume that that maxim, or at least the overall Cooperative Principle, is observed at the level of what is implicated" (Grice, 1975, p. 52). Quality is seemingly broken in irony to imply a meaning other than what is explicitly said — "the most obviously related proposition is the contradictory of the one he purports to be putting forward" (Grice, 1975, p. 53). For this to work, some intelligibility is necessary to intuitively understand that what is said is not necessarily what is implied, because understanding irony assumes that the CP is still in play. As will be further discussed, current algorithmic regulations struggle to differentiate these subtleties of meaning, negatively impacting agency and discursive democracy.

### Discursive democracy and agency: Habermas' lifeworld and public sphere; how does misinformation in social media affect our agency?

In this section, we consider the question of individual and collective agency in datafied, algorithm-driven online environments in light of Jürgen Habermas' discourse-centered theory of democracy. We see how, for Habermas, agency and autonomy are tied to communicative interactions and democratic institutions. Building on Habermas' concepts of the lifeworld and public sphere, we discuss how the transforming conditions of communication may affect participation in deliberative democracy. We begin to reflect on the justifications for regulations concerning online misinformation.

Applying critical theory, it has been Habermas' intention to provide a "diagnosis of the times" (Fultner, 2014, p. 8), and one of the central themes in Habermas' work concerns communicative practices. In Habermas' research, the intention has been to trace and interpret the socio-cultural norms and structures governing the practices and institutions especially in the cultural hemisphere of Western Europe (Anderson, 2014, pp. 91–94; Froomkin, 2003, p. 760). Following on the previous section's account of the Cooperative Principle and the vitality of truth-telling to communicative interaction, we now focus on democracy as the institutionalization of the implicit ideals of communication, such as equality, openness and inclusion (Habermas, 1992, p. 448; Olson, 2014, p. 144). While citizens themselves might not be aware of these ideals, according to Habermas' theoretical framework they nevertheless work as a backdrop to our communicative actions (Olson, 2014, p. 141). "This discourse-centered concept of democracy places its faith in the political mobilization and utilization of communicative force of production," Habermas (1992, p. 447) states in "Further Reflections on the Public Sphere", a postscript to his germane *The Structural Transformation of the Public Sphere* (1962). Habermas (1992, p. 447) continues, "social issues liable to generate conflicts are open to rational regulation, that is, regulation in the common interest of all parties involved" (see also Froomkin, 2003, p. 766; Olson, 2014, p. 144). The question of self-determination is, in Habermas' description, central to democracy:

> [Democracy's] essence consists of the fact that it enacts far-reaching social changes that increase the freedom of human beings — and ultimately, can perhaps create them in the first place. Democracy works upon mankind's self-determination, and only when the former is real [*wirklich*] is the latter true [*wahr*]. Political participation is identical to self-determination. (Habermas, J. 1973. *Kultur und Kritik: Verstreute Aufsätze*. Frankfurt: Suhrkamp, p. 11; as cited in Schmalz-Bruns, 2017, p. 123; original emphasis)

Democracy is generative of individuals as citizens and persons because who we are and how express ourselves, and how we relate to each other, are functions of the political system we were raised in.

In the words of Joel Anderson (2014, p. 92), Habermas is "a staunch defender" of individual freedom; however, Habermas rejects individualistic philosophy of consciousness and the traditional empiricist view of the sovereign self-governing individual. While autonomy and agency are often viewed as individual concerns, in Habermas' thinking, they are fundamentally social constructs: private autonomy does not exist without public autonomy. Together, they form political autonomy. In his account of Habermas' approach on political autonomy, Anderson draws attention to this "dual emphasis" on the individual and the social, describing how political autonomy emerges together with social practices, political institutions and legal rights. Anderson writes, "[it] is not that autonomy becomes difficult without the framework of rights or the institutions of democratic decision-making; it ceases to exist" (Anderson, pp. 96, 91–96).

The understanding of autonomy as a social construct is linked to Habermas' conceptualization of the *lifeworld*, the intersubjective web of language and culture within which we make the world intelligible to us. The communicative infrastructure of the lifeworld is "always already" there (Habermas, 1987, p. 119), providing context for our social interaction. We, the participants, continually reproduce the lifeworld via communicative actions and thus provide more resources for communication. We use, test and renew cultural tradition; both the lifeworld and we as participants evolve and are limited by the transformations that take place in society (Habermas, 1987, pp. 119–126, 131, 137–139; Gilbert, 2018, p. 89). For self-determination and autonomy to prevail, to an extent a rational lifeworld is required (Anderson, 2014, p. 105). The public sphere, as described in the previous section, represents the part of lifeworld wherein (ideally) self-determined individuals engage in debate and form public opinion, directed towards consensus (Garnham, 1992, pp. 358–359). According to Habermas (1992, p. 453), a functioning public sphere needs, in addition to the constitutional institutions, "the supportive spirit of cultural traditions and patterns of socialization, of the political culture, of a populace accustomed to freedom."

While much of our interaction is symbolic in nature, it is very much dependent on the material and economic conditions of communication. In today's world, as we increasingly make meaning of the world and communicate in online environments, this question of the material dimension of lifeworld is of utmost importance. Both online and traditional media can influence our agency and self-governance in beneficial or disadvantageous ways; what is beyond dispute is that technology companies exercise great and often opaque power over our social resources (Fuchs, 2014, pp. 57–58; Gilbert, 2018, pp. 89, 92). Andrew Simon Gilbert (2018) analyses how, in light of critical theory, the use of algorithms transforms the nature of communication and reflects on the effects of data-driven feedback loops on democracy and culture. Gilbert (2018, pp. 90–92) brings up concerns regarding the privatization of communication and notes that online, the content "has always already been filtered through and organized by computerized processes"; sometimes it is even produced by social bots (Shao et al, 2018). When our intersubjective communication is mediated by systems that operate according to functional, nonlinguistic logics, the danger arises that, in Habermasian terms, *systems colonize the lifeworld*: our democratic culture depends on us as human participants engaging in communication, aimed at achieving agreement, but "this is prevented when instrumental systems have already determined our decisions for us" (Gilbert, 2018, p. 89; see also Amadae 2018a, pp. 20–21).

When the lifeworld is reproduced instrumentally and algorithms have control over what we are exposed to, they often selectively magnify certain content while leaving something else out (Gilbert, 2018, pp. 91–92). Recent research by Shao et al. (2018) also shows that manipulative content and misinformation may spread more easily online. They note: "[w]hile fabricated news are not a new phenomenon, the ease with which social media can be manipulated creates novel challenges and particularly fertile grounds for sowing disinformation" (Shao et al., 2018, p. 2), thus skewing our "always already" there infrastructure of social interaction. Our vulnerability lies in our intersubjective dependence of communicative actions, and this form of life requires protection (Anderson, 2014, p. 91).

Alongside colonization, nevertheless, exists the possibility for *recolonization* (see Gilbert, 2018, p. 92). Christian Fuchs (2014, p. 89) writes, "social media has a potential to be a public sphere and lifeworld of communicative action, but that this sphere is limited by the steering media of political power and money so that corporations own and control and the state monitors users' data on social media". Fuchs (2014, p. 97) calls for democratic reforms of social media in the name of public interest, echoing Habermas' (1987, p. 444, original emphasis) own words that demand us "to erect a democratic dam against the colonializing *encroachment* of system imperatives on areas of the lifeworld". In the following sections, as we proceed to review the German Network Enforcement Act and the Nordic Model that represent efforts to enhance democratic self-determination online, we will also consider the role of algorithms in this democratization.

## Case Study: "NetzDG" and Traditional Gatekeepers

Following the Second World War, Germany passed numerous laws against Holocaust denial and more generally with regards to inciting hatred against "groups determined by nationality, race, religion or ethnic origin" (Haupt, 2006, p. 323). Today, unified Germany maintains some of the world's toughest laws regarding hate speech (Hawdon, Oksanen, & Räsänen, 2016, p. 5), and there has been continual concern among politicians in the country about a lack of online accountability towards these legal standards.

To shape regulation of online platforms towards this same direction, the *Network Enforcement Act* (*Netzwerkdurchsetzungsgesetz* or *NetzDG*) introduced in Germany in 2017, presents a set of strict reporting guidelines regarding illegal content for online platforms with more than 2 million users located in the country. Once reported, platforms must investigate the content, and if it is found to be "obviously unlawful" it must be removed within 24 hours, with other illegal materials required to be taken down within seven days, at the threat of fines up to €50 million (Bundesrepublik Deutschland, 2017). Promisingly, the legislation has codified accountability and transparency in the form of regular public reporting from platforms such as Facebook, YouTube and Twitter. Still, there remain large unanswered questions regarding "freedom of expression and the potential chilling effects of legislation" (Tworek & Leerssen, 2019, p. 1) as well as the potentially troubling model it provides for more authoritarian regimes.

NetzDG has been commonly referred to as a "hate speech law", but with its targeting of large online platforms and implementation of the law against rising misinformation, it has also been labelled as the "Facebook act" by the media. Despite some overarching criticism which we will examine next, it is very notable that NetzDG does not define new categories of illegal content specifically for the web. Instead, the law is designed to extend and enforce 22 existing criminal code statutes on social platforms and hold these companies directly responsible for their continued enforcement (Tworek & Leerssen, 2019, p. 2). Besides traditional hate speech, these statues also target offenses that resemble the phenomenon of fake news via statues targeting *intentional defamation*, *treasonous forgery*, and *forgery of data* (Claussen, 2018), with actions against these types of information already being taken under the law (Twitter, Inc., 2019). This type of application of existing statutes to the Internet would seem in line with popular proposals of those critical of online misinformation to extend traditional media and defamation legislation towards the Internet. The presence of reintroducing gatekeepers in combating egregious speech online would certainly seem worthwhile to limit the reach of those that would espouse hate speech, but the tradeoff in this case is difficult to fully justify. As Hawdon, Oksanen, & Räsänen (2016, p. 8) found in the pre-NetzDG Internet of 2016, German nationals were already at the lowest risk of seeing or being exposed to hate material among comparative democracies. When they compare this against a country like the United States, they find a 55% majority of their sample being regularly exposed to online hate material.

The bill, while drawing criticism during its drafting both internally and internationally, seems largely popular with the German public, where one poll shows 67%

“strongly approving” of the law (Deveaux, 2018). One of the most distinct outside criticisms for the law came from the UN Special Rapporteur on Freedom of Opinion, David Kaye, who in a statement to the German government found the provisions of the Act incompatible with international human rights declarations such as the International Covenant on Civil and Political Rights. He strongly criticized the implementation of in the draft legislation as an endangerment to human rights, stating that “while it is recognized that business enterprises also have a responsibility to respect human rights, censorship measures should not be delegated to private entities” (Kaye, 2017, p. 2). This responsibility on the online platforms empowers their algorithms to delete content upon “vague and ambiguous criteria” and would lead to inappropriate interference in freedom of expression and privacy; as Kaye outlines: “liability placed upon private companies to remove third party content absent a judicial oversight is not compatible with international human rights law” (Kaye, 2017, p. 4).

Although German law constitutionally guarantees freedom of speech and the press within limits, this pursuit of legislators towards strict hate speech legislation has often had the effect of reducing freedoms. In the case of NetzDG, there is an increasing fear of these unintended effects when applying the legal statues to the online space with this broad-brush algorithmic approach. That fear, and the responsibility for “privatized enforcement” shifting entirely to these multinational corporations, has caused a real need for rethinking this type of legislation in the context of promoting deliberative democracy. The responsibility of social platforms has commonly had the effect of producing an algorithmic policy of “delete-in-doubt” by the social media giants, producing an amount of over-removal in their compliance to NetzDG (Kinstler, 2018).

One of the most critical concerns now with the type of content moderation being so widely adopted in Germany is how it might serve as a blueprint, inspiration or justification for authoritarian regimes around the world to restrict speech (Tworek & Leerssen, 2019, pp. 2–4). With the news of the Russian State Duma already using the NetzDG as a model for similar legislation, Christian Mihr, managing director of Reporters Without Borders (Germany) painted a dark picture of the future this law is inspiring:

> Our worst fears come true: The German law against hate messages on the Internet now serves undemocratic states as a template to restrict social debates on the Internet. […] From now on in Russia, social networks will be forced to decide under time pressure which information will be deleted. In a country without independent courts that could enforce the protection of freedom of expression, this is a devastating development (Reporter ohne Grenzen e.V, 2019).

The implementation of legislation, inspired by NetzDG, but without an independent judiciary is worth reflecting upon. Within states already implementing stricter controls over the Internet, this legislation can be observed to have a very damaging effect within the Quality maxim of the Cooperative Principle. For example, with the algorithm effectively being blind to the use of irony inside of online discourse, removal of content of this nature has a serious implication for the intelligibility required for truth-telling in communications. Algorithms do not have an intelligible grasp of context, and yet apply judgments ostensibly requiring and intelligible grasp of the lifeworld in which meanings are crafted and conveyed. In more extreme cases, prosecution and punishment against online speech that exhibits irony can be observed, creating a fundamental question about human rights to free speech.

Even holding the most cynical view of modern news corporations, one can point to a fundamental difference between the increasingly consolidated tech giants of the Internet age and the new and old media companies and publishers that previously held a near monopoly on the flow of mass media content in television, radio, newspapers and books (Gilbert 2018, p. 92). With companies more and more wanting to act as the neutral and ubiquitous platform — e.g. Facebook would seem to be comfortable imagining themselves *as* the Internet instead of an actor in a larger system — it is much harder to hold them to the standards and self-imposed regulations of publishers and other historical gatekeepers. Professional editors that make a career and may pride themselves on being gatekeepers upholding Habermas' ideals of debate and discussion are not the same as the algorithm and employees that fulfill the compliance of these platforms to legislation like NetzDG and its offsprings. Thus prior to the gigantic new social media and Internet companies, citizens as members of civil society could participate in the state's government, as well as in the myriad democratically constituted municipal governments and even the constitution of private organizations. Yet increasingly multinational corporations host and govern the platforms on which civil discourse is now performed.

## Discussion

With the concept of the *Cooperative Principle* by Paul Grice, and Jürgen Habermas' conception of deliberation in the public sphere, the previous sections provided the basic parameters for our article's view on ideal communication in the digital sphere. The practical focus on the German *NetzDG* as a regulative attempt for online hate speech offered the one exemplary attempt of implementing a censorship regime under the premise of strengthening democratic and pluralist debate, while ensuring the sanctioning of prosecutable content. The following section continues along these lines of thought by first providing a theoretical perspective on the shortcomings of algorithm-based online censorship, their potential threat to deliberative democracy, and potential alternative interpretations of algorithms' usefulness as administrational tools. In its second part, the section takes a more practical standpoint, looking at the *Nordic Model* as a possible alternative to external regulation.

### Algorithmic paradigms

Here we take a closer look into inner functioning and shortcomings of algorithms. We will point out risks against the principles of deliberative democracy stemming from the reliance on the algorithmic tool in online moderation. We offer a partial solution in terms of a suggested theoretical approach through a paradigmatic shift.

### Intrinsic obstacles of the algorithm

The algorithmic ability for self-restructuring in reaction to the exposure to data is often referred to as "machine learning". This terminology overlooks that these differences contrasted with a human conception of *learning* are vast since "machines learn by finding statistical relationships in past data" (Boulanin, 2019, p. 16). The AI's exposure to data has an ultimate and direct influence on its "development", making this term possibly more adequate than the term "learning" with its implications of deliberate reflexivity. Based on the inner working of the machine learning, which is dependent upon its training set, and lacks any contextual ability to frame data beyond its training set, Vincent Boulanin (2019, 16, pp. 18–21) points out the central shortcomings of the current state of AI.

The reliance on AI self-development solves practical problems by reducing the need for intensive hand coding and makes it possible to reach software complexity beyond human competence (Boulanin, 2019, pp. 15–16). However, the data exposure's effects on the algorithmic compositions developed through machine learning can hardly be foreseen by its engineers. The engineers' ability to merely observe rough inputs and outputs to and from an otherwise "black box" demonstrates the central problem of *algorithmic opacity* which results in essential predictability problems. This weakness is not fully inscribed in the machine but is rather a reinforcement of human failure at providing adequate "training data." As AI lacks any understanding of what humans would describe as "common sense" — or comprehensive situational awareness consistent with being members of a lifeworld — it cannot only deliver misrepresentations, but also reinforce human biases (e.g. algorithmic racial bias). "If the training data set is not representative, then the system might fail, might perform poorly, or might misinform human decisions and actions by reinforcing existing human biases or creating new ones" (Boulanin, 2019, p. 19). Thinking about the assumption that algorithms are expected to deliver a neutral, optimal judgement as their output, we must realize that they are rather characterized by their performance mirroring human imperfection.

### Archiving the future — Challenges for deliberative democracy

The wide-ranging contemporary application of algorithms and their central shortcomings in interpretative tasks do not seem aligned, and might be explainable by a societal desire for simple, actionable truths. Eventually, it would be an orientation informed by *computable rationality* towards achieving overarching utility maximization (Amadae, 2018b, pp. 193–197) that generates the need for those optimized numeric outputs.

As the prior case study showed, it is not necessarily the situation that the state hands its oversight of sanctionable elements in the public debate immediately over to algorithmic automation. Rather, one may recognize the manifestation of digital corporations' *structural power* (Horten, 2016) in controlling online communication by the handover of the task of content selection to machine intelligence. Again, referring to Gilbert (2018, pp. 90, 92), it is the combination of "privatization of communication in a political-economic sense" (Gilbert, 2018, p. 90). Privatization's consequential control over the communicative (online) space with its cultural and semantic resources that results in the *colonization of the lifeworld* (Habermas, 1987). This process is realized through algorithmic content moderation as a chain of delegation (from state to corporation) and efficiency maximization (from corporation to the algorithmic tool). Traditionally it has been the liberal-democratic constitutional state's objective to only persecute cases of criminal relevance — and especially leave the moderation of public debate to the considerations of intelligible human agents (e.g. journalists, authors etc.). However, now it is the privately imposed algorithmic content governance that has turned the long hope for dialogic media communication into a structure of hollow predictive content generalization and normalization (Brecht, 2000 [1932]; Harper, 2017, pp. 1428, 1436).

A limitation of the scope of tasks and capabilities of algorithms applies in a technical sense. But much more importantly, it is vital to separate between this *capability* and the *effective use* of algorithms, thereby highlighting the importance of the differentiation between technical capability and its human application. One and the same tool may be able to track customers in the retail industry and perform military target identification. The determining component in a normative sense of regulating these processes lies in human hands. Yet, a persistent concern is the possible abandoning of human agency in favor of artificial decision making. This concern respects either the process of advising human agents which are then at risk of being structurally led to rely too heavily on the algorithmic 'recommendation' (e.g. in the criminal justice system; cf. Kehl, Guo, & Kessler, 2017) or as the partially autonomous entity in a decision process (e.g. the AI behind online censorship).

It is the basic task and capability of algorithmic sensemaking to reduce a multiplicity of possibilities to an *optimized* output (Amoore, 2019). However, what seems to be generally misinterpreted in terms of the expectations of the technology is the general belief in algorithms to deliver an *optimal* output in the sense of neutral and rational calculation. This positivistic understanding cannot only be challenged through the argument of the definitory prevalence of the human decision, as it was explained above.

As for the algorithmic tools, we can summarize two main challenges to deliberative processes. First, the algorithm's incapacity to identify patterns of *conversational implicature* like irony (Grice, 1975, pp. 50–53), but also to distinguish cultural-historical values, poses the threat of "over blocking". Therefore the spectrum of public debate becomes limited, and the machine learned interpretation of significance serves to define what is normal. These trends threaten to create a form of "communication that denies intention, meaning, and intelligibility" (Amadae, 2018b, p. 189).

Secondly — and adding to the problem of normalization — the issue of machine "learning" with its essential goal of behavioral and processual *prediction* (Boulanin, 2019, pp. 16, 18–21), is to be seen as having an immediate effect on public discourse. This can be seen through the adaption of subjects' behavior (e.g. getting accustomed or adapting one's communication to the algorithmic incapacity to 'understand' irony) or through the foreclosure of alternative ways of action through the dominance of the optimized algorithmic output. Louise Amoore (2018, pp. 16–20) phrases this issue as an "archiving of the future." The AI's capability to detect patterns in past data is used to condense "particular future connections […] from the volume of the data stream […], opening the possibility for seemingly infinite calculability" (Amoore, 2018, p. 16), even though it handles its data samples detached from their context. Opposing a societal belief in a paradigm that also characterizes human agents through computational rationality (Amadae, 2018b, pp. 190–191), the harm is not in giving control from a human agent to the machine per se. Rather, the key problem lies in not acknowledging the uncertainty which is inscribed into the algorithmically optimized outputs as they reject multiplicities of alternatives in favor of one numeric recommendation (Amoore, 2019). Louise Amoore (2018, pp. 12–13; 2019) therefore suggests to rather think of the algorithmic tool as an *aperture*, that immediately illustrates the technique of choosing an optimized output out of a multiplicity of available alternative options. In other words, the more that digital communication technologies rely on algorithms and machine learning to optimize outputs and make judgments, the more we may foreclose on a future open with imaginative possibilities of meaning generated by inhabiting our lifeworld. Instead of co-constructing our future prospects, we end up dwelling in an "AI-world" structured by computationally derived symbolic outputs archived from our past and served up by machines trained on confined data sets.

## What can smart regulations look like? The case for agency and trust enhancing initiatives in fighting misinformation

After looking at the practicality of the German example of external online content regulation and the theoretical underpinning of algorithmic tools, it is worth looking at what smart regulations could look like. External content control does not necessarily support truth telling or encourage good communication methods. Instead, its punitive character penalizes those who lie by removing their content from platforms. The problem is that it might only provide a temporary fix for the long-term problem of misinformation. Furthermore, it does not support individual agency but decides for the individual user what shall be deemed to be true or false. It does not trust consumers' ability to identify fake news, and neither does it teach them how to possibly identify false information. This further reiterates the unsustainability of this type of regulation and poses the question of what smart regulation that encourages individual agency and enhance truth telling online could look like.

One direction to explore is the *Nordic Model*. This report, published in 2018 by the Nordic Council of Ministers, does not set out content regulation per se, but rather a vision of a system of mutual trust between the recipient of news and the traditional media that would empower the individual to not be manipulated by possible misinformation

(Bjerregård & Lundgren, 2018). It constitutes a dual approach that focuses both on the elevation of content provided by traditional primary information providers, and the empowerment of the individual to identify misinformation. *The Nordic Model* sees what they call *true news* as a weapon against fake news emphasizing the importance of trust within society ("Copenhagen experts meeting reflects on 'fake news'", 2018, p. 11). These authors therefore must understand the diminishing degree of trust in true news to be a reason for the belief in misinformation of the *alternative media.* To reestablish this trust, they set out certain guidelines for true news that promote more precision, fairness, and transparency in reporting (Hanson, 2018, p. 16). Furthermore, the report commits itself to a more pluralistic media culture ("Copenhagen Experts Meeting Reflects on 'Fake News'", 2018, p. 11), next to public crowd checking (Hanson, 2018, pp. 14–15) that encourages the readerships' critical engagement with the information they receive. Additionally, it sets out to make media and information literacy part of school curricula to not only enhance individuals' ability to identify false information but also to support a healthy debate culture (Weihe, 2018, p. 28). The Nordic approach compared to the German approach exemplifies an emphasis on self, rather than external, regulation. It is hereby worth looking at the main philosophical underpinnings of this approach, that is the enhancement of trust and the emphasis on self-regulation, and why these underpinnings are important for the encouragement of individual agency and truth telling.

Trust is an important sociological concept that both influences social relations (Cook, Hardin & Levi, 2009, p. 88) and the context within which individuals receive and interpret news (Sterret et.al., 2019, p. 2). On the role of trust in deliberative democracies Mark E. Warren (1999) states in his book *Democracy and Trust*, that democratic regimes need stability and a general culture of trust (Warren, 1999, p. 7). The type of mutual understanding that is facilitated by trust is vital in deliberative democracies to come to a conclusion of political issues that require deliberation (Warren, 1999, p. 18). Trust is an essential pillar for any civic culture that aims to enhance solidarity and cooperation. When trust is low, citizens may be worried that their interests are not taken into consideration and feel misrepresented (Cook, Hardin & Levi, 2009, p. 310). Looking at communication as cooperative exchange (Amadae, 2018a, p. 19), trust that the communicative partner is telling the truth is just as essential as the effort of the individuals' commitment to speak the truth. It would thus not matter whether the communicative actor believes oneself to be speaking the truth when the other does not trust what one says to be true: the goal of communication will not be fulfilled. Enhancing trust in the information that citizens consume and communicate is therefore essential, in order to return to a mutual understanding of truthfulness which crucial for the cooperative principle to be valid. The Nordic Model emphasizes the traditional, *true media*, to be essential for a mutual understanding of truth, seeing the enhancement of trust in traditional media as the key to fight misinformation and the disagreement over what truth is.

The centrality of traditional media as a political actor and the main provider of information is defined by the symbiotic relationship between citizens, politicians, and the media in democracies (Broersma & Peters, 2013). Marcel Broersma (2013) sees the claim to truth as the core of journalism stating that "As a producer of knowledge, journalism derives its authority from its presumed ability to provide a truthful representation of the social world" (Broersma, 2013, p. 31). The function of a truthful provider of information in a democracy therefore is crucial in enabling individuals to act as citizens (Broersma, 2013, p. 31). Social media however has altered the relationship between the sender and receiver of information (Broersma, 2013, p. 15) blurring the line between public and private sphere. This notion supports the *Nordic Model's* belief in the importance of the preservation of the role of the media as the main provider of *true news* that is able to be regulated and commits itself to certain standards of truth telling. The guidelines for a higher quality of *true news* emphasize the centrality of traditional media as an actor in the fight against fake news (Bjerregård & Lundgren, 2018).

Next to the goal of increasing trust in traditional media, the *Nordic Model* also focuses on enhancing individual agency by suggesting greater citizen involvement with traditional media through public readership crowd checking (Hanson, 2018, pp. 14–15) as well as efforts to enhance media and information literacy through its inclusion in school curriculums (Weihe, 2018, p. 28). This is part of an effort to avoid external regulation by supporting self-regulation through the strengthening of individual agency (Bjerregård & Lundgren, 2018, p. 39). The importance on literacy education and the fostering of debate culture in the classroom is supported by many scholars who see these as vital in fighting misinformation online (Shao et al., 2018, p. 2; Delacruz, 2009, p. 14). Evidence for the success of literacy education that encourages critical thinking can be seen in Finland, which has the highest media literacy rate and resilience to misinformation in the world, a result of extensive literacy education (Mackintosh, 2019). Within deliberative democracies the support of individual agency regarding the issue of misinformation is important for individuals to be able to participate effectively in deliberative processes on public goods on the basis of truthful information for the best possible outcome. This would not happen if individuals were not be able to distinguish between truthful and false information, making the support of individual agency essential for upholding deliberative democracy in online communication.

The dual approach of the Nordic Model does offer a valid alternative to the German model and could provide a more long-term approach to the fight against online misinformation. Yet, one must keep in mind that for the Nordic Model to work, traditional media have to be committed to being truthful, impartial, and independent actors. Social bots remain a key player in the spread of misinformation (Shao et al., 2018, p. 5), which is highly problematic. A combined approach of algorithmic external control to stop the speed of spread of misinformation and Nordic Model-like initiatives to stop it from being misidentified as truthful could prove to be a more effective option.

## Conclusion and Outlook

This paper's aim was to analyze the role of different types of government initiatives in the fight against misinformation, based on Habermas' deliberative democracy paradigm. It focused on the role of Grice's Cooperative Principle and of agency in the public sphere and the surrounding lifeworld as our communication increasingly takes place online. The section on Grice's Cooperative Principle emphasizes the importance of truth-telling in communication and highlights the importance of intelligibility to identify the true meaning of what is said.

The next section evaluated the role of agency in Habermas' conception of discursive democracy. The presentation of Germany's *NetzDG* law demonstrated what privatized algorithmic control of online content could look like. Its transference of responsibility to identify misinformation from the user to social media platforms was shown to have ramifications for the freedom of speech, open Internet, and its possible adaptation by authoritarian regimes. After discussing this practical case, we developed an abstract perspective on the transference of state oversight to relying on algorithmic tools for administrating online communication. We contrasted the algorithm's intrinsic shortcomings as exemplifying human imperfection to societal expectations of easily actionable (numeric) truth. We pointed out the challenges of this expectation and an unreflected use of the algorithm to modes of deliberative democracy, and suggested a paradigm change in our acceptance of the algorithmic tools as a means of action. Most importantly, algorithmic control prioritizes prediction over action, meaning that it prevents alternative outcomes from occurring, and thus possibly preventing meaningful civic discourse to take place. Relying on algorithmic control does not support individual media literacy capabilities, which leaves consumers more vulnerable when misinformation is mishandled by algorithmic governance.

The Nordic Model has served as an example of a government initiative that does not solely rely on regulation by algorithms, but looks inward to society to solve the problem of misinformation. It focuses on the elevation of content provided by traditional media actors to increase civic trust. The Nordic Model therefore emphasizes trust in these actors to have a central role in fighting misinformation. The section demonstrated that trust plays an important role in deliberative democracy by referring to the Cooperative Principle. Additionally, the Nordic Model reiterates the importance of media literacy education to enhance individual agency in the fight against misinformation. The Nordic Model proves to be not only a valid alternative to external control in the fight against misinformation but also a more sustainable solution. By enhancing essential concepts such as the Cooperative Principle and individual agency, it complements Habermas' paradigm of deliberative democracy. Yet, one must keep in mind that the model can also be idealistic and may only work in societies in which traditional media actors are committed to the truth and education is independent of politics. Furthermore, the Nordic Model cannot prevent the spread of misinformation through, for example, social bots and can only provide citizens with tools to evaluate misinformation as incredible. External regulation that is aimed at curbing social bots combined with the Nordic Model could present an even more effective tool in the fight against both the spread and manipulation of digital misinformation.

## References

Amadae, S. M. (2018a). Game Theory, Cheap, and Post-Truth Politics: David Lewis vs. John Searle on Reasons for Truth-Telling. *Journal for the Theory of Social Behavior*, *48*(3), pp. 1–24. https://doi.org/10.1111/jtsb.12169

Amadae, S. M. (2018b). Computable Rationality, NUTS, and the Nuclear Leviathan. In D. Bessner & N. Guilhot (Eds.), *The Decisionist Imagination: Sovereignty, Social Science and Democracy in the 20th Century* (pp. 173–214). New York: Berghahn.

Amoore, L. (2018). Cloud Geographies: Computing, Data, Sovereignty. *Progress in Human Geography*, *42*(1), 4–24.

Amoore, L. (2019). Our Lives with Algorithms. Alexander von Humboldt Institute for Internet and Society. Retrieved from https://www.hiig.de/events/louise-amoore-our-lives-with-algorithms/

Anderson, J. (2014). Autonomy, Agency, and the Self, in B. Fultner (Ed.), *Jürgen Habermas: Key Concepts*. Routledge, pp. 90–111. ProQuest Ebook Central, https://ebookcentral-proquest-com.libproxy.helsinki.fi/lib/helsinki-ebooks/detail.action?docID=1886909.

Bjerregård and Lundgren (2018) Recommendations, in Bjerregård, M. B. & Lundgren, P. (Eds.),

*Fighting Fakes – The Nordic Way. Nordic Council of Ministers.* Retrieved October 24,

2019, from https://www.nordicom.gu.se/en/latest/news/fighting-fakes-nordic-way

Boulanin, V. (2019). Artificial Intelligence: A Primer. In V. Boulanin (Ed.), *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk. Volume I: Euro-Atlantic Perspectives* (pp. 13–25). Stockholm: Stockholm International Peace Research Institute.

Braaten, J. (1991). *Habermas' Critical Theory of Society.* Albany: State University of New York Press.

Brecht, B. (2000). Der Rundfunk als Kommunikationsapparat [The Radio as an Apparatus of Communication]. In C. Pias, J. Vogl, L. Engell, O. Fahle, & B. Neitzel (Eds.), *Kursbuch Medienkultur. Die maßgeblichen Theorien von Brecht bis Baudrillard* (pp. 259–263). Stuttgart.

Broersma, M. J. & Peters, C. (2013). *Rethinking journalism: Trust and participation in a transformed news landscape.* New York: Routledge.

Bundesrepublik Deutschland. *Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken [The Network Enforcement Act].* (2017).

Claussen, V. (2018). Fighting Hate Speech and Fake News. The Network Enforcement Act (NetzDG) in Germany in the context of European legislation. *Rivista Di Diritto Dei Media*, *2018*(3), pp. 110–136. Retrieved from http://www.medialaws.eu/wp-content/uploads/2019/05/6.-Claussen.pdf

Cook, K. S., Hardin, R. & Levi, M. (2009). *Whom can we Trust? How Groups, Networks, and Institutions Make Trust Possible*. New York: Russell Sage Foundation.

"Copenhagen Experts Meeting Reflects on 'Fake News'" (2018), in Bjerregård, M. B. & Lundgren, P. (Eds.), Fighting Fakes – The Nordic Way. Nordic Council of Ministers. Retrieved October 24, 2019, from https://www.nordicom.gu.se/en/latest/news/fighting-fakes-nordic-way

Delacruz, E. M. (2008). From bricks and mortar to the public sphere in cyberspace: Creating a culture of caring on the digital global commons. *International Journal of Education & the Arts*, *10*(5). pp. 1–21. Retrieved October, 28, 2019, from http://www.ijea.org/v10n5/.

Deveaux, F. (2018, April 17). 87% of Germans Approve of Social Media Regulation Law – Dalia Research. Retrieved October 30, 2019, from Dalia Research website: https://daliaresearch.com/blog-germans-approve-of-social-media-regulation-law/

Sterrett, D., Malato, D., Benz, J., Kantor, L., Tompson, T., Rosenstiel, T., . . . Loker, K. (2019). Who Shared It?: Deciding What News to Trust on Social Media. *Digital Journalism, 7*(6), pp. 783–801. https://doi.org10.1080/21670811.2019.1623702

Froomkin, A. M. (2003). Habermas@Discourse.Net: Toward a Critical Theory of Cyberspace. *Harvard Law Review*, *116*(3), pp. 749–873. https://doi.org/10.2307/1342583

Fuchs, C. (2014). Social Media and the Public Sphere. *tripleC: Communication, Capitalism & Critique*, *12*(1), pp. 57–101. https://doi.org/10.31269/triplec.v12i1.552

Fultner, B. (2014). Introduction, in B. Fultner (Ed.), *Jürgen Habermas: Key Concepts.* Routledge, pp. 1–12. ProQuest Ebook Central, https://ebookcentral-proquest-com.libproxy.helsinki.fi/lib/helsinki-ebooks/detail.action?docID=1886909.

Garnham, N. (1992): The Media and the Public Sphere, in C. Calhoun (Ed.), *Habermas and the Public Sphere.* Cambridge, MA: MIT Press, pp. 359–376.

Gilbert, A. S. (2018). Algorithmic Culture and the Colonization of Life-Worlds. *Thesis Eleven*, *146*(1), 87–96. https://doi.org/10.1177/0725513618776699

Grandy, R.E. and Warner, R. (Winter 2017 Edition). "Paul Grice". In Edward N. Zalta (Ed.). *The Stanford Encyclopedia of Philosophy*. Retrieved October 24, 2019, from https://plato.stanford.edu/archives/win2017/entries/grice/.

Grice, H. P. (1975). Logic and Conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and Semantics 3: Speech Acts* (pp. 41–58). New York: Academic Press.

Grice, H. P. (1991). *Studies in the way of words.* Cambridge, Mass.: Harvard University Press.

Habermas, J. (1984). *The Theory of Communicative Action*, *Vol. 1.* (T. McCarthy, Trans.). Boston: Beacon Press.

Habermas, J. (1987). *The Theory of Communicative Action, Volume 2: Lifeworld and Systems: A Critique of Functionalist Reason*. Cambridge: Polity Press.

Habermas, J., (1991). *The Structural Transformation of the Public Sphere: An Inquiry into a Category of Bourgeois Society*. (T. Burger and F. Lawrence, Trans.). Cambridge, MA: MIT Press.

Habermas, J., (1992). Further Reflections on the Public Sphere, in C. Calhoun (Ed.), *Habermas and the Public Sphere*. Cambridge, MA: MIT Press, pp. 421–461.

Hanson (2018). True news against fake news, in Bjerregård, M. B. & Lundgren, P. (Eds.), *Fighting Fakes – The Nordic Way. Nordic Council of Ministers*. Retrieved October 24,

2019, from https://www.nordicom.gu.se/en/latest/news/fighting-fakes-nordic-way

Harper, T. (2017). The Big Data Public and its Problems: Big Data and the Structural Transformation of the Public Sphere. *New Media & Society*, *19*(9), pp. 1424–1439.

Haupt, C. E. (2006). Regulating Hate Speech- Damned If You Do and Damned If You Don't: Lessons Learned From Comparing the German and U.S. Approaches. *Boston University International Law Journal*, *23*(2), pp. 299–333.

Hawdon, J., Oksanen, A., & Räsänen, P. (2016). Exposure to Online Hate in Four Nations: A Cross-National Consideration. *Deviant Behavior*, *38*(3), pp. 254–266. https://doi.org/10.1080/01639625.2016.1196985

Horten, M. (2016). *The Closing of the Net*. Cambridge: Polity Press.

Kaye, D. (2017). *Mandate of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression*. Retrieved from https://www.ohchr.org/Documents/Issues/Opinion/Legislation/OL-DEU-1-2017.pdf

Kehl, D., Guo, P., & Kessler, S. (2017). Algorithms in the Criminal Justice System: Assessing the Use of Risk Assessments in Sentencing. Responsive Communities Initiative, Berkman Klein Center for Internet & Society, Harvard Law School.

Kinstler, L. (2018, May 18). Germany's Attempt to Fix Facebook Is Backfiring. Retrieved October 31, 2019, from The Atlantic website: https://www.theatlantic.com/international/archive/2018/05/germany-facebook-afd/560435/

Mackintosh, E. (2019, May). *Finland is winning the war on fake news. Other nations want the blueprint*. Retrieved October 31, 2019, from https://edition.cnn.com/interactive/2019/05/europe/finland-fake-news-intl/.

Olson, K. (2014). Deliberative Democracy, in B. Fultner (Ed.), *Jürgen Habermas: Key Concepts*. Routledge, pp. 140–155. ProQuest Ebook Central, https://ebookcentral-proquest-com.libproxy.helsinki.fi/lib/helsinki-ebooks/detail.action?docID=1886909.

Reporter ohne Grenzen e.V. (2019). Russland kopiert Gesetz gegen Hassbotschaften [Russia copies law against messages of hate]. Retrieved October 30, 2019, from Reporter ohne Grenzen für Informationsfreiheit website: https://www.reporter-ohne-grenzen.de/pressemitteilungen/meldung/russland-kopiert-gesetz-gegen-hassbotschaften/

Schmalz-Bruns, R. (2017). The Theory of Democracy, in Brunkhorst, H., Kreide, R., & Lafont, C. (Eds.), *The Habermas Handbook* (Vol. English-language edition). New York: Columbia University Press, pp. 123–132.

Shao, C., Ciampaglia, G. L., Varol, O., Yang, K., Flammini, A., & Menczer, F. (2018). The Spread of Low-Credibility Content by Social Bots. *Nature communications, 9*(1), pp. 1–9. https://doi.org/10.1038/s41467-018-06930-7

Twitter, Inc. (2019). *Network Enforcement Act, Report January – June, 2019*. Retrieved from Twitter, Inc. website: https://transparency.twitter.com/en/countries/de.html

Tworek, H., & Leerssen, P. (2019). *An Analysis of Germany's NetzDG Law*. Retrieved from https://www.ivir.nl/publicaties/download/NetzDG_Tworek_Leerssen_April_2019.pdf

Warren, M. E. (1999). *Democracy and trust*. Cambridge: Cambridge University Press.

Weihe (2018). Democracy first, in Bjerregård, M. B. & Lundgren, P. (Eds.), *Fighting Fakes –*

*The Nordic Way. Nordic Council of Ministers.* Retrieved October 24, 2019, from https://www.nordicom.gu.se/en/latest/news/fighting-fakes-nordic-way

# 6.2 The Reactions of Political and Corporate Authorities to the Impact of the Information Revolution

Veera Villikari, Alina Mäkynen, Helmi Hämäläinen, Heli Hämäläinen,
Aleksander Heikkinen
Faculty of Social Sciences, University of Helsinki

## Abstract

The information revolution and the innovations accompanying it provide many opportunities and challenges to the ways in which power is distributed and exercised in societies. This has led to different reactions from corporate and political authorities, and the aim of this research paper is to analyze five of them. While the EU has tried to regulate the power of private data companies, China has tried to benefit from the data by developing a social credit system. Meanwhile, the Cambridge Analytica scandal in the context of the 2016 US presidential elections demonstrates a "warning sign" of the misuse of personal data. In addition, two influential media corporations, Facebook and Twitter, are coping with misinformation, election interference and questions of privacy. These cases demonstrate the possible impacts of the information revolution and provide possible strategies to cope with its opportunities and challenges.

*Keywords*: Big data, Facebook, Cambridge Analytica, Twitter, China, Social Credit System, General Data Protection Regulation, GDPR, Twitter, Facebook, fake news

The development and welfare of societies are increasingly dependent on information technologies. They are reshaping our personal identities by affecting our social life, economy and decision-making (Floridi, 2014.). The wide adoption of new technologies has prompted many concerns, including the protection of our privacy. Moreover, information technologies are creating opportunities for false information to spread through social media platforms such as Facebook and Twitter. These platforms can be used for propaganda purposes and clickbaiting with political or economic objectives (Benkler, Benkler, Faris, & Roberts, 2018, pp. 9–12). The current social environment has created friction on political decision-making and legal measures to tackle it (Floridi, 2014, p. 102).

Technologies as such have not led the change. They are developed and adopted in certain cultural, political and institutional environments. The same technologies have different effects in political systems with different historical backgrounds and traditions (Benkler & al., 2018, p. 9). A case in point is presented in this research paper. While the EU is focusing on the threats of personal data processing, China is developing a credit system in which big data technologies are utilized.

The aim of this research paper is to evaluate how different political and corporate authorities have reacted to the impacts of the information revolution. The research will concentrate on five practical cases. The examples have been selected based on the significance of their impact on society. The first section introduces the case of Facebook and its responsibility and possible actions to counter the spread of fake news and information on its platform. The second section proceeds to evaluate Twitter's complex role in the public sphere. The discussion then continues to the case of Cambridge Analytica and its practice of harvesting and misappropriating personal data for political purposes. The research paper then moves on to observe the political impact of the information revolution and continues to discuss its role in data protection in the EU. The last case study introduces China's social credit system which is built on people's historical and ongoing economic and social activities. Finally, the conclusion summarizes the main findings of our research

## Facebook, algorithms and the spread of false information

Social networking platforms' reach is so widespread and information is transmitted at such a fast pace that distorted, inaccurate and false information has the potential to reach millions of users rapidly. This has raised concerns about individuals' ability to recognize and evaluate misleading content. Detecting false news and clickbait has become a central discussion on Facebook and demands for the company to take responsibility for the content and advertisement posted on their platform have been increasing. In addition, Facebook has received criticism over its role in the 2016 US presidential election for allowing the false news industry to use the platform as a seedbed for spreading propaganda and fake news. A large portion of pro-Trump content was traced to Macedonia, where teenagers with little interest in American politics had figured out that by imitating actual news sites their publications received

Facebook engagements from Trump supporters, which translated into advertising dollars. Facebook also partnered up with private companies such as Acxiom and allowed campaigns to target its users in hyper-specific segments created from their personal data (Benkler, Faris, & Roberts, 2018).

### The spread of false news

Since fake news is a widely used term, it is necessary to specify how it is used in this research paper. A group of researchers (Shu, Sliva, Wang, Tang, & Liu, 2017) defined false news as the type of news articles that are intentionally and verifiably false, and were conceived with the goal of misleading readers. The researchers emphasized two key elements of this definition: authenticity and intent. False news includes false information and can be verified as such, and the false information is created with the intention of misleading consumers. To be wholly precise with terminology, the phrase "fake news" draws attention to its counterfeit nature: this genre of communication mimics authentic news sources but spreads false information.

To make a distinction between authentic and false news, reporters of *The Guardian* have reported on a parliamentary inquiry into fake news to consider legislation on editorial responsibility for social media companies and to force them to take responsibility. The reporters stated that Facebook is the principal advertisement tool for political communication and should be held accountable for its content. Damian Collins, from the UK's culture, media and sport (CMS) committee admitted that democracy might be compromised in the future by the high level of virality of fake news (Gupta, Pangotra, Prahbat, & Bajaj, 2017).

Fake news tends to spread fast and gain large audiences in a short period of time. To understand the spread of fake news Nabiha Syed introduces a theory of amplification (2017). According to Nabiha the amplification principle explains the cycles of misinformation through filters and permeating communities. The theory consists of two stages: first the provocative ideas percolate in remote corners of the Internet and second, the ideas find their way into the mainstream media. This is important for two reasons: it reveals how fragile and prone to manipulation online information filters are, and these information filters shape perception of what is true and what is false. Individuals are more likely to find often repeated and familiar statements true. The Facebook algorithm makes sure that individuals who are already ideologically inclined to believe false news will continue to be less interested in the truth. This cycle makes the individuals see the like-minded articles but not the debunking theories.

### Facebook's response to fake news

Facebook publicly announced that the amount of fake news on its platform is such a small percentage that it could not have had an impact on the US election. At the same time, the company has officially insisted that distinguishing between authentic news and false news is a difficult technical problem (Figueira & Oliveira, 2017). Interestingly, it only took 36 hours for a few Princeton students to create a Facebook browser extension that detects false news and unreliable sources. The algorithm checks the user's personal news feed and labels news items according to the system's verification (Gupta et al., 2017).

Facebook's co-founder Mark Zuckerberg has commented on censorship on Facebook several times. He has given controversial statements on Facebook's responsibility over the content shared on the platform. Zuckerberg has argued that "Facebook is in the business of letting people share stuff they are interested in" (Syed, 2017). After the 2015 Charlie Hedbo attacks in Paris, Zuckerberg stated that Facebook does not intend to censor itself; and after 2016, Zuckerberg explained that Facebook is a platform "for all ideas" (ibid). Facebook has claimed to deprioritize links that are shared by suspected spammers. They are launching new features to recognize clickbait, sensationalism, and misinformation and to make users think twice before sharing a story. Facebook has also banned verified false news websites from gaining revenue by using its advertising program. These actions may be useful for filtering out profit-oriented false news content, such as the Macedonian teenagers, but not for stopping propaganda and misinformation that is fueled mainly by political interests.

In the most recent statement published on Facebook's blog for the social networking site, the Vice President of Facebook, Adam Mosseri, voiced the company's concern over the spread of misinformation and false news. Mosseri introduces Facebook's efforts to stop misinformation. Their objectives include: disrupting economic incentives, building new products to stop the spread of false news, helping people to make more informed decisions and recognize false news. He proceeds to state that Facebook is not in a position to "become arbiters of truth" (Mosseri, 2017). This is in line with Zuckerberg's previous position that Facebook does not plan to censor the content shared by Facebook's users.

Mosseri's blog post introduces Facebook's aim to improve News Feed ranking by determining that if users do not share an article, it is more likely to be false information than if the they do (Mosseri, 2017). This statement conflicts with the study results of Pennycook and Rand (2019), who found that sharing intentions are not driven by the user's perception of the content's accuracy. According to their study, users share content if it supports their reputation, regardless of the information's accuracy.

## Discussion

A solution for the spread of misleading information, spam, and fake news needs to be found. The use of algorithms and machine learning are helpful tools in finding fake news in a user's feed. After the verification of the News Feed, it is left to the user to evaluate what to do with the flagged content. If the verification process is done exclusively by algorithmic recognition, different types of content such as humorous and sarcastic publications need to be taken into consideration. The quest for a system to prevent and censor false news may also collide with democratic values such as freedom of speech.

Facebook's representatives seem to give vague statements with regard to the company's actions for fighting fake news. According to a study conducted by Pennycook and Rand (2019) the possibilities for developing filtering tools which can be adjusted to the Facebook News Feed exist. Mosseri and Zuckerberg have made it clear that their concerns are more likely to be in the area of avoiding the violation freedom of speech than addressing the ongoing debate on the spread of fake news.

## How does Twitter regulate the public sphere?

One of the biggest concerns in the era of the information revolution is whether new media platforms create a pseudo public sphere where irrational ways of thinking become dominant. New technologies have changed political communication and challenged democracy. Technologies are not just platforms; they also shape thinking and are the means of political actions (Finlayson, 2019, pp. 77–79).

In this case study we will look at Twitter which has achieved a position as a mainstream medium of political communication. Moreover, it influences public opinion. There are several reasons why Twitter has an important and visible role in society, even if it is not the most widely used social media platform (Isotalus, Jussila, & Matikainen, 2018, p. 9). First, Twitter allows information to spread fast since sharing information does not require following the original poster (Colleoni, Rozza, & Arvidsson 2014, p. 319). Second, the political elite and journalists have adapted the usage of Twitter widely (Isotalus et al., 2018, p. 9).

One of the major problems with regard to Twitter is the spread of digital misinformation which is claimed to influence elections and threatens to undermine democracies (Shao et al., 2018, p. 2). Twitter is one of the largest online platforms, which has been called on to deploy preventive measures and algorithms against election interference and the spread of fake news. The question of state actors using media platforms for political interference is a major one. Bots, which are software-controlled profiles, present another problem as they deliver misinformation to those who are most likely to believe it (ibid.) and likely do not realize this content is automatically generated.

### Twitter in the public sphere

The criticism social media companies face is at the heart of contemporary political communication that has changed rapidly (Finlayson 2019, p. 80). Colleoni, Rozza and Arvidsson (2014) have studied Twitter's role as a medium of political communication and its impact on the public sphere. According to them, the public sphere is a communicative space wherein valid information can be circulated in an unfettered way which contributes to the formation of political will via deliberation (Colleoni, 2014, p. 318; Dahlgren, 2005, p. 148). Twitter has the potential to either increase the diversity of opinions or to function as an echo chamber. In the latter case, people's preexisting perspectives are reinforced which channels the formation of public opinion (Colleoni, Rozza, & Arvidsson, 2014, p. 317).

Colleoni et al. (2014, 318) state that the mediums of communication do not reinforce democratic conversation, they only strengthen already dominant political views due to the selectiveness of the exposure increasing the heterogeneity of political discussion. From this point of view Twitter can function as an echo chamber, because individuals with similar views form ties between each other. Colleoni et al. (2014, p. 319) explain this phenomenon through the theories of cognitive dissonance and selective exposure suggesting that people tend to seek information that confirms their opinions. This creates homogeneous groups and when applied to the political domain, it can result in polarization.

Research indicates that Twitter can be conducive toward deliberation in the public sphere when it is analyzed and treated as a news medium. On the other hand, when looked as a social medium and permitted to operate as such, it can be seen more as an echo chamber (ibid., 2014, p. 328).

### What is the response?

In March 2018 Twitter introduced an approach aimed to improve the public conversation. There is not much information about this project, but in their blog post Twitter's product managers claimed that one of their biggest challenges has been the eradication of disruptive behavior that does not violate Twitter's policies per se, but has a negative impact on public discourse (Gasca & Harvey, 2018).

In that post, the managers describe the methods Twitter uses to filter content. They state that signals that convey disruptive behavior are not usually visible. They give a few examples of such signals such as "an account has not confirmed their email", "if the same person signs up for multiple accounts simultaneously", "accounts that repeatedly Tweet and mention accounts that don't follow them", and "behavior that might indicate a coordinated attack" (Gasca & Harvey, 2018). They also reveal that they monitor how accounts interact with those that violate these rules. According to Twitter's representatives, violations were eight percent lower in the conversations in which this feature was tested. Twitter also recently reported (Dorsey, 2019) their new plan to ban political advertisement. The announcement will be confirmed later, but it already raises concerns about how the ads are classified as political, and the possible non-transparency of the vetting algorithms.

Even if Twitter's own algorithms are not transparent. Several studies have suggested ways to enhance the public sphere of Twitter. Shao et al. (2018, p. 2) propose that if the media platforms create echo chambers, new algorithms should be used to broaden the users' exposure to diverse information. Furthermore, bot-driven abuse can be detected by machine learning algorithms. Shao et al. (ibid., p. 5) also propose that curbing social bots may be effective in mitigating the spread of false information.

Oxford Internet Institute's Computational propaganda project has also developed a computational method to identify deliberative manipulation. By using the Coefficient of Traffic Manipulation (CTM) method, Twitter feeds can be ranked as organic or manipulated. Twitter traffic can be manipulated by a small group of users who generate a large flow of traffic, disproportionate to the number of users involved. The manipulation can be done by automated bot accounts, partially-automated accounts or human-run accounts (Nimmo, 2019, pp. 5–7). For example, one of the study's cases looked at Twitter traffic about Marine Le Pen's campaign during the French presidential election. In this case, the traffic was created by a high-volume combination of human users and bots (ibid., p. 13). Nimmo (2019, p. 19) believes CTM to be a useful first warning of Twitter traffic that might be manipulated.

### Discussion

As is often the case, big data is controlled and owned by private companies whose algorithms are not transparent. When decisions affecting the public are formed via commercial platforms, a conflict in the public sphere increases. It is important to note that ideally the public sphere is either fair or free (Finlayson, 2019, p. 80).

On one hand Twitter can be used as a tool to spread fake information and conspiratorial thinking, which runs the risk to geopolitical destabilization. On the other hand, it can offer a platform for those who are otherwise excluded from political participation. What needs to be done is to develop new strategies and rhetoric to cope with the challenges and possibilities. Changes in the digital sphere pose the question of whether Twitter usage can be organized in a way that enhances democratic freedom. (Ibid., pp. 78–80.)

As a media company with enormous power, Twitter should introduce more effective ways of regulating their users, as well as new ways to protect victims of harassment. Despite the fact that social media platforms are already acknowledging the problems mentioned, and tend to deploy countermeasures, it is hard to evaluate whether these measures are effective (Shao et al., 2018, p. 5). However, in the light of recent studies on Twitter, there are many effective methods which Twitter could use to enhance the regulation of its content.

## When nothing is done: Cambridge Analytica

In March 2018 yet another scandal relating to president Trump's 2016 election campaign broke out. Christopher Wylie, an ex-employee of the data company Cambridge Analytica, revealed that the company had harvested the data of about 87 million people's Facebook profiles and used it for political advertising purposes — without the users' consent (Kozlowska, 2018). The notorious scandal provides a dystopian example of how the digital revolution's innovations — as in this case big data through data harvesting — can in an unregulated context lead to privacy violations and misuse of personal data for purposes such as political microtargeting.

### How it was done

Cambridge Analytica's vast database was collected through the Facebook app "thisisyourdigitallife". Hundreds of thousands of Facebook users, who agreed to have their data collected for academic use, took a personality test that revealed, among other things, their names, locations and Facebook "likes". What the test-takers didn't realize was that the app simultaneously collected the same information from their Facebook friends, resulting in a database of around 87 million — over a quarter of potential US voters. The acquired data was used to build an algorithm that could analyze individual Facebook profiles and determine personality traits linked to voting behavior. This sort of information constitutes a powerful political tool, as it enables the identification of possible swing voters and the creation of microtargeted political messages that are more likely to resonate (Cadwalladr & Graham-Harrison, 2018). Cambridge Analytica's own executives claimed openly that they were able to carry the Electoral College in Trump's favor in 2016 by manipulating only 40,000 voters in three states. Although the company's claim of influence on the result of the election might be over-exaggerated, it is nevertheless clear that they did manage to play a major role in the election (Berghel, 2018, pp. 85–86). Additionally, the company was at the time headed by Trump's key adviser and chief executive of his presidential campaign, Steve Bannon, who also later became the Chief Strategist and member of the US National Security Council for the Trump administration (Cadwalladr & Graham-Harrison, 2018).

Cambridge Analytica's parent company, the London-based Strategic Communications Laboratories (SCL) was introduced to the concept of using social media data for political profiling by a Cambridge University's lecturer, Aleksandr Kogan. Kogan, who is the founder and director of Global Science Research (GSR), began working with SCL to deliver a "large research project in the US" in which SCL agreed to pay for GSR's data collection in order to, among other things, enhance GSR's algorithm's "national capacity to profile capacity of American citizens" (Davies, 2015) SCL has also been revealed to work as a defense contractor for governments and militaries around the world, with electoral influence in developing countries playing an instrumental part (Cadwalladr, 2019). Kogan on the other hand has been discovered to have received funding from the Russian government, as well as from a hedge fund millionaire and leading Republican donor Robert Mercer — who also happenned to be the owner of Cambridge Analytica (Cadwalladr & Graham-Harrison, 2018). Mercer also owns the intellectual property (IP) of Aggregate IQ, which is a web analytics company that worked as one of the major forces in delivering Brexit for the official Leave campaign in Britain. Although Aggregate IQ has officially declined having any formal relationship with Cambridge Analytica, it was found to have a revolving cast of data scientists who went on to work with Cambridge Analytica and vice versa (Cadwalladr, 2017).

### The responses

The 2018-scandal was in fact not the first time Cambridge Analytica's illicit activities were exposed. Already in 2015, *The Guardian* released an article on how the campaign of Republican Senator Ted Cruz had paid the company to collect psychological profiles of potential voters. After the revelation, Facebook declared that they would carefully investigate the situation, and that misuse of their information is a direct violation of their policies as well as a breach of trust (Davies, 2015).

Nevertheless, in the end Facebook failed to alert users and only took limited steps to recover and secure their private information (Cadwalladr & Graham-Harrison, 2018).

The 2018 case however turned out different. The story made headlines all over the world and Facebook's share price plunged more than 50 billion dollars (and subsequently continued to fall over twice of that) (Cadwalladr, 2019). While Zuckerberg refused to answer the questions about Facebook's role in the scandal, the official statement was that Facebook was not guilty of a data breach as its systems were not penetrated, and the data was instead mishandled by a third-party against Facebook's terms of service.

However, because customers' experience was that their trust had been broken, Zuckerberg issued official full-page apologies in nine major US and international newspapers, and declared that as an act of self-regulation Facebook will voluntarily applicate the European General Data Protection Regulation (GDPR) to all Facebook customers (and not just EU citizens) (Kozlowska, 2018). The governments' response was also more severe this time around: Britain's information commissioner obtained a warrant to enter Cambridge Analytica's offices and seize its servers, and Mark Zuckerberg was called in to be questioned both in the US Congress and the European Parliament (Cadwalladr, 2019). The President of the European Parliament, Antonio Tajani, called the case an unacceptable violation of EU citizens' privacy rights, and promised an EU investigation (Manokha, 2018, p. 892). Elizabeth Warren, the Democrat candidate for 2020 presidential race even called for the breaking up of Facebook (Cadwalladr, 2019). Meanwhile Cambridge Analytica and SCL announced their closure in May 2018 (Manokha, 2018, p. 892).

## Discussion

Considering that Facebook simultaneously functioned as the source of the information and a platform on which the created content was delivered on a large scale, it is not difficult to understand the importance of government regulation when dealing with big data companies. User data has become a special kind of commodity that is collected, processed, analyzed and finally monetized in one way or another (Manokha, 2018, p. 902). This is evident in the way the business model of several data companies, such as Facebook, relies on their access to their users' data for targeted advertising (Kozlowska, 2018). This results in a major threat to the right of privacy, which in turn can compromise the exercise of several other forms of individual freedom, such as freedom of expression, thought, and press (Manokha, 2018, p. 903). Besides these obvious troubling implications to people's privacy, the use of big data for political purposes arguably also poses a threat to democratic systems as a whole. As in this case, data harvesting was conducted for the purpose of achieving efficient microtargeting, which is a "pseudo-public" form of communication. This means that the personalized content is not under the scrutiny and deliberation in the "marketplace of ideas", where they could be exposed and challenged. Thus, it also makes it easier to share misinformation and blurs the idea between advertising and other forms of content, resulting in the threat of voters' pure manipulation (Heawood, 2018, pp. 432–433.) Meanwhile in the eyes of the political campaigns, votes are commodified; the goal is to manipulate and predict rather than understand the voters' views, values, needs and desires (Gorton, 2016, p. 73).

## The European Union – a role model in data protection?

One of the present and future political challenges is the creation of realistic regulatory standards for the use of data and algorithmic decision-making (Mittelstadt, Allo, Taddeo, Wachter, & Floridi, 2016, p. 13). The European Union seems to have taken a special interest in the protection of personal data which is currently one of its core values (Brkan, 2016, p. 813). Its objective at data protection appears in the General Data Protection Regulation (GDPR) that took effect on May 2018 (EC, 2016).

### The architecture of the GDPR

Data protection has received a lot of public attention in recent years (Brkan, 2016, p. 813). Yet it is not a new phenomenon. The EU has had a Data Protection Directive since 1995 which set minimum standards for personal data protection inside the EU (Eliantonio, Galli, & Schaper, 2016, p. 392). However, the platforms and the volume in which data are collected and used have changed rapidly. Roughly a decade ago actors, such as the EU officials and NGOs, began a debate on the old directive being outdated (Starkbaum & Felt, 2019, p. 5) in the age of information technology where the flow of personal data is increasing constantly (Eliantonio et al., 2016, p. 398). Controversies of data processing were often associated with US companies, such as Facebook, that had been in the center of attention in the processing of personal data (Krystlik, 2017, p. 6)[1]. Moreover, the EU lacked consistency in data protection within the member states. This was considered detrimental for the internal market and businesses, where data processing was becoming increasingly important (Eliantonio et al., 2016, p. 399).

In the summary of the GDPR, the European Commission states that the aim of the regulation is to give EU citizens more control over their personal data and to unify data protection rules across the member states (EC, 2016.). The regulation defines personal data as "any information relating to" data subject, "such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person" (Art. 4(1) GDPR, 2016). In addition to the protection of individuals, the EU hopes that harmonized rules will reduce bureaucracy and therefore benefit companies. The impact of GDPR extends to non-EU companies, as they have to adopt the rules when they process data of individuals that are located in the EU (EC, 2016).

One core legal basis provided by the GDPR is consent of individuals (data subjects) in processing. The regulation specifies consent as "any freely given,

[1] For example, in 2013 activist Max Schrems raised questions about the use of European personal data in his campaign against Facebook. The case was covered widely in the media. (Krystlik, 2017, p. 6.)

specific, informed and unambiguous indication of the data subject's wishes by which he or she, by a statement or by a clear affirmative action, signifies agreement to the processing of personal data relating to him or her" (Art. 4(11) GDPR, 2016). Further, data subjects are allowed to withdraw consent anytime (Botta & Wiedemann, 2019, p. 431).

## Effectiveness of the regulation

GDPR has attracted a lot of attention before and after its adoption. The media framed it as groundbreaking because its legal validity harmonized national rules in Europe (Starkbaum & Felt, 2019, p. 6). But even as scholars (Botta & Wiedemann, 2019; Ferretti, 2018; Kamleitner & Mitchell, 2019; Mittelstadt et al., 2016) recognize the GDPR's potentials in protecting individuals' data, they underline the possible loopholes and inefficiencies in its adoption. This section focuses on the data subjects' perspective leaving out the shortcomings encountered by data controllers, such as companies and public officials.

Botta and Wiedemann (2019, pp. 431–432) argue that the whole ideology behind GDPR is to give data subjects full knowledge and control over their data. They note that subjects are assumed to act fully informed when they make decisions. However, data subjects might not have the willingness or ability to decide freely whom to give consent. GDPR fails its objective when the owners of personal data do not act self-determinedly and autonomously. In addition, data processors know the potential limitations and might take advantage of the situation (Ibid).

In agreement with Botta and Wiedemann (2019), Ferretti (2018) argues that one of the main shortcomings of GDPR is "the consent of consumers as the legitimising ground" (p. 499). He bases the argument on his analysis on how GDPR responds to the challenges of FinTech and big data from consumers' perspective. First, the complexities of, for example, "FinTech business models", and "data-collection practices" may create a misunderstanding of the possible consequences of data processing. Moreover, the relationship between consumers and vendors might be imbalanced. If a consumer needs credit she has to accept the terms, as the lack of consent to data processing could stop the company from giving credit. (Ibid., p. 496.)

Mittelstadt et al. (2016, pp. 13–14) consider the GDPR as an example of the difficulties that political actors encounter in regulating algorithms. From data subjects' perspective, GDPR seems to enable individuals to make decisions with regard to their data. However, they argue that the regulation contains exemptions concerning the rights of individuals as data subjects. For example, rights can be restricted if data controllers have legitimate reason for processing which "override the interests, rights and freedoms of the data subject" (Ibid., p. 17).

One exemption that was debated during the preparation of GDPR is that data subjects are allowed to "give a one-time consent for their data to be used for multiple scientific research projects across time" (Starkbaum & Felt, 2019, p. 6). Supporters of strict data protection (e.g. the European Parliament in the beginning of preparation) wanted to ban data reuse if a new consent was not given. The debate demonstrates how information society is seeking to find a balance between individual rights and the desired collective benefits in data processing (Ibid., pp. 1–3). According to Brkan (2016), the EU emphasizes data protection which often prevails over other interests such as "public security, freedom of information and economic interests of the controller" (p. 814). However, the above-mentioned exemption indicates that data protection does not prevail the idea of collective benefit produced by science.

Lastly, Kamleitner and Mitchell (2019, p. 435) argue that GDPR neglects to protect personal data as it assumes that the original owner is identifiable. For example, when data subjects agree to conditions in an app, they might give consent to the use of their personal data, such as pictures or contacts, that they don't own to begin with. This means that if a data subject decides to withdraw consent from a company, other individuals can still share that person's information through their consent. This creates pressure for the EU to decide if data subjects have the right to share data only if they own it (Ibid., pp. 435, 446).

### Discussion

The GDPR is still a new piece of regulation and we are only starting to see its effects. It aims to give EU citizens more control over their personal data by encouraging citizens to decide who has the rights to use it. Nevertheless, the GDPR struggles with unsolved issues, such as the ambiguity in data ownership and asymmetrical relationships between companies and individuals.

The protection the GDPR will provide depends on its legal interpretations: the framework can be either powerful or weak from the data subjects' perspective (Mittelstadt et al. 2016, p. 14). Eliantonio et al. (2016, pp. 402–403) emphasize the influence of national and European courts that have always played a central role in balancing the conflicting interests of data protection. Overall, GDPR appears to be a necessary endeavour of legislative framework aiming at the protection of data. However, it seems to be more indicative of the difficulties that political authorities face in data regulation than a robust legislation.

## China's Social Credit System – can data points generate trust?

Big data technologies have enabled and expanded the array of surveillance capacities that nation states have at their disposal. One prominent use of these technologies can be found in China, where the government seeks to create a nationwide Social Credit System (SCS) that scores the "trustworthiness" and "creditworthiness" of 1.4 billion people by 2020 (Liang, Das, Kostyuk, & Hussain, 2018). This system is based on data about people's historical and ongoing economic and social activities and will determine their rights in society through rewards and punishments (Creemers, 2018; Liang et al., 2018). Thus, the SCS is used as a tool to monitor, manage and predict the behaviour of individuals and businesses with the intention of enforcing trust in society (Creemers, 2018; Liang et al., 2018). This program is characterized by its entanglement with the extensive use of big data technologies, as it requires massive amounts of data to be collected, stored, shared and analyzed (Chen & Cheung, 2017).

## Social Credit System – what is it about?

The SCS is not (yet) a single system but consists of multiple fragmented initiatives that are developed either by local governments or business entities. These different programs all aim at influencing the behaviour of individuals and businesses by sanctioning or rewarding them (Kostka, 2019). These different initiatives can be separated into schemes that either seek to determine the "creditworthiness" or "trustworthiness" of different actors in society (Creemers, 2018). The core of the latter schemes is a joint punishment system that consist of red lists and black lists. Individuals who comply with the set standards end up on red lists which allow them to enjoy certain perks (e.g. faster commuting), whereas misbehaving citizens are placed on black lists, which means they are cut-off from several public goods, such as public transportation (Creemers, 2018; Liang et al., 2018.). This punishment system is often binary, meaning that people are either on or off the black list (Creemers, 2018).

In turn, the financial credit system that seeks to determine "creditworthiness" of citizens is technically more advanced. That is, these programs often include an actual score that effects individuals' opportunities, such as one's possibilities of getting a loan (Creemers, 2018). So far these commercial SCS initiatives resemble and function like traditional "loyalty schemes" (Kostka, 2019). Nevertheless, the Chinese government seeks to develop a nationwide SCS that incorporates the functions of these initiatives into a single SCS (Creemers, 2018; Kostka, 2019).

## The development of credit schemes

These efforts of expanding traditional credit ratings into evaluating social behaviours more broadly are widely portrayed as petrifying by Western media. In contrast to these views, the SCS has gained high levels of approval in China. As an example, 80 percent of respondents in one study either strongly approved or somewhat approved of the SCS, and less than 2 percent strongly disapproved or somewhat disapproved of the system (Kostka, 2019). These differences in interpretations can be largely explained by the misconceptions that Western media has on the subject (Creemers, 2018; Kostka, 2019; Liang et al., 2018). Thus, in order to understand the Social Credit System, it is vital to acknowledge the context in which it has evolved.

The first considerations towards the SCS emerged in the 1990's in discussions related to the modernization of China's market economy. These discussions pinpointed trust as a critical element of a prosperous market and consequently led the Chinese government to seek ways to improve the financial creditworthiness and conducts of honesty and trust in the marketplace. This project took a large step forward in 2007, when concrete policy measures were established in order to construct a SCS (Creemers, 2018).

Afterwards, between 2007 and 2013 these considerations spawned minor breakthroughs towards the SCS, for example in the form of local initiatives and early trials of different IT infrastructures. These developments eventually culminated in 2014 when a concrete plan was launched to implement a nationwide social credit system by 2020. (Creemers, 2018.)

The SCS is widely welcomed in Chinese society by its citizens and government, both of whom are frustrated with the nation's moral decline and people's inability to trust one another (Creemers, 2018; Kostka, 2019). Indeed, according to some scholars (e.g. Creemers, 2018; Kostka, 2019), the SCS should not be perceived merely as an "Orwellian nightmare" as it is often described. Instead, they suggest that the program should be seen as a prominent tool in decreasing the prevailing dishonesty in society and thus improving the lives of Chinese citizens[2]. Opposing views (e.g. Chen & Cheung, 2017; Hoffman, 2017; Liang et al., 2018) largely consider the SCS as a tool for authoritarian–like social management.

### Big data-enabled Social Credit System

Setting aside the different conceptualizations of the purposes of the SCS, a common ground can be found when looking at the connection between SCS and the utilization of digital era technologies. Indeed, both sides of the argument seem to accept the technological novelty that the SCS aim to utilize (Chen & Cheung, 2017; Creemers, 2018; Hoffman, 2017; Kostka, 2019; Liang et al., 2018).

The utilization and development of these high–end technologies were outlined in the 2014 SCS roadmap, alongside with the blueprints for the bureaucratic and financial support systems (Creemers, 2018). The nationwide reach and complexity of the SCS requires excessive amounts of data to be collected, which has led the Chinese government to invest in big data–enabled technologies in order to collect, mine and analyze data. Consequently, China has instrumentalized and institutionalized big data innovations and information communication technologies (ICTs) (Liang et al., 2018).

Liang et al. (2018) suggests that the data collection and integration of the SCS can be understood to work in three phases: data collection, data aggregation and data analytics. In the first phase, data is collected from various public and private sources (e.g. bank statements, criminal records, social media use). Secondly, the data is aggregated from separate platforms by sharing and integrating the data for social ratings. This data is then evaluated in the final phase, leading to decisions on whether subjects are placed on a red or black list.

### Discussion

China has rapidly informatized its governance in order to transform the way it manages both the state and society (Creemers, 2018). Although this transformation is still ongoing, the SCS has the potential of radically transforming the state's governance in new directions, notably through the utilization of big data technologies. Some scholars consider the direction of this development as a way of integrating trust in society, whereas others provide more or less hyperbolic claims about the emergence of an "Orwellian"–like dystopia.

How the SCS will develop largely depends on how the data points are computed into the social score. Therefore, the transparency of the computation is

[2] Kostka's (2019, p. 1585) survey findings suggest that Chinese citizens consider the SCS as an instrument to improve "quality of life", rather than an instrument of "surveillance".

essential; if the algorithms that determine people's scores are ambiguous and opaque, the SCS will most likely resemble a system of oppression. On the other hand, if the scores are processed in a transparent manner, the SCS can reduce moral decline and enhance trustworthiness in Chinese society.

## References

Benkler, Y., Faris, R. and Roberts, H. (2018). *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics.* New York: Oxford University Press.

Berghel, H. (2018). Malice domestic: The Cambridge Analytica Dystopia. *Computer* 51(5), 84–89.

Botta, M., & Wiedemann, K. (2019). The interaction of EU competition, consumer, and data protection law in the digital economy: the regulatory dilemma in the Facebook odyssey. *The Antitrust Bulletin*, 64(3), 428-446.

Brkan, M. (2016). The Unstoppable Expansion of the EU Fundamental Right to Data Protection: Little Shop of Horrors?. *Maastricht Journal of European and Comparative Law*, 23(5), 812-841.

Cadwalladr, C. (2017). The Great British Brexit robbery: how our democracy was hijacked. *The Guardian*. Available at: https://www.theguardian.com/technology/2017/may/07/the-great-british-brexit-robbery-hijacked-democracy [Accessed 20 October 2019].

Cadwalladr, C. (2019). Cambridge Analytica a year on: 'a lesson in institutional failure'. *The Guardian*. Available at: https://www.theguardian.com/uk-news/2019/mar/17/cambridge-analytica-year-on-lesson-in-institutional-failure-christopher-wylie [Accessed 20 October 2019].

Cadwalladr, C. & Graham-Harrison, E. (2018). Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. *The Guardian*. Available at: https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election [Accessed 17 October 2019].

Chen, Y., & Cheung, A. S. Y. (2017). The transparent self under big data profiling: privacy and Chinese legislation on the social credit system. *The Journal of Comparative Law*, 12(2), 356–378.

Colleoni, E., Rozza, A. & Arvidsson, A. (2014). Echo Chamber or Public Sphere? Predicting Political Orientation and Measuring Political Homophily in Twitter Using Big Data. *Journal of Communication* 64(2), 317–322.

Creemers, R. (2018). China's Social Credit System: An Evolving Practice of Control. *SSRN Electronic Journal*.

Dahlgren, P. (2005). The Internet, public spheres, and political communication: Dispersion and deliberation. *Political Communication*, 22(2), 147–162.

Davies, H. (2015). Ted Cruz using firm that harvested data on millions of unwitting Facebook users. *The Guardian*. Available at: https://www.theguardian.com/us-news/2015/dec/11/senator-ted-cruz-president-campaign-facebook-user-data [Accessed 16 October 2019].

Dorsey, J. [Jack]. (2019, October, 30). We've made the decision to stop all political advertising on Twitter globally. We believe political message reach should be earned, not bought. Why? A few reasons… [Tweet] Available at: https://twitter.com/jack/status/1189634360472829952 [Accessed 4 November 2019].

Eliantonio, M., Galli, F., & Schaper, M. (2016). A balanced data protection in the EU: Conflicts and possible solutions. *Maastricht Journal of European and Comparative Law*, 23(3), 391–403.

European Commission. (2016). Summary of Regulation (EU) 2016/679 — protection of natural persons with regard to the processing of personal data and the free movement of such data. Available at: https://eur-lex.europa.eu/legal-content/EN/LSU/?uri=uriserv:OJ.L_.2016.119.01.0001.01.ENG [Accessed 30 October 2019].

Ferretti, F. (2018). Consumer access to capital in the age of FinTech and big data: The limits of EU law. *Maastricht Journal of European and Comparative Law*, 25(4), 476¬499.

Figueira, Á., & Oliveira, L. (2017). The current state of fake news: challenges and opportunities. *Procedia Computer Science*, 121, 817-825.

Finlayson, A. (2019). 7. Rethinking Political Communication. *The Political Quarterly* 90(S1), 77-91.

Floridi, L. (2014). *The Fourth Revolution: How the Infosphere is Reshaping Human Reality.* Oxford: Oxford University Press.

Gasca, D. & Harvey, D. (2018). Serving healthy conversation. [Blog] Twitter. Available at: https://blog.twitter.com/official/en_us/topics/product/2018/Serving_Healthy_Conversation.html [Accessed 20 October 2019].

General Data Protection Regulation 2016. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679 [Accessed 30 October 2019].

Gorton, W. A. (2016). Manipulating Citizens: How Political Campaigns' Use of Behavioral Social Science Harms Democracy. *New Political Science* 38(1) 61–80.

Gupta, A., Prabhat, P., Gupta, R., Pangotra, S., & Bajaj, S. (2017, December). Message Authentication System for Mobile Messaging Applications. In 2017 International Conference on Next Generation Computing and Information Systems (ICNGCIS) (pp. 147-152). IEEE.

Heawood, J. (2018). Pseudo-public political speech: Democratic implications of the Cambridge Analytica scandal. *Information Polity* 23(4) 429–434.

Hoffman, S. R. (2017). *Programming China: The Communist Party's autonomic approach to managing state security* (Doctoral dissertation, University of Nottingham).

Isotalus, P., Jussila, J. & Matikainen, J. (2018). Twitter viestintänä ja sosiaalisen median ilmiönä. In Isotalus, P., Jussila, J., Matikainen, J. and Boedeker, M. (ed.), *Twitter Viestintänä: Ilmiöt Ja Verkostot*. Tampere: Vastapaino, 9–31.

Kamleitner, B., & Mitchell, V. (2019). Your data is my data: A framework for addressing interdependent privacy infringements. *Journal of Public Policy & Marketing*, 38(4), 433-450.

Kostka, G. (2019). China's social credit systems and public opinion: Explaining high levels of approval. *New Media & Society*, 21(7), 1565–1593.

Kozlowska, I. (2018). Facebook and Data Privacy in the Age of Cambridge Analytica. University of Washington. Available at: https://jsis.washington.edu/news/facebook-data-privacy-age-cambridge-analytica/ [Accessed October 21, 2019].

Krystlik, J. (2017). With GDPR, preparation is everything. *Computer Fraud & Security*, 2017(6), 5-8.

Kostka, G. (2019). China's social credit systems and public opinion: Explaining high levels of approval. *New Media & Society*, 21(7), 1565–1593.

Liang, F., Das, V., Kostyuk, N., & Hussain, M. M. (2018). Constructing a Data ‐ Driven Society: China's Social Credit System as a State Surveillance Infrastructure. *Policy & Internet*, 10(4), 415 ‐ 453.

Manokha, I. (2018). Surveillance: The DNA of Platform Capital – The Case of Cambridge Analytica Put into Perspective. *Theory & Event* 21(4) 891–913.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2), 1–21.

Mosseri, A. (2017). Working to stop misinformation and false news. Newsroom.fb.com.

Nimmo, B. (2019). Measuring traffic manipulation on Twitter. Computational Propaganda Research Project, Oxford Internet Institute. Available at: https://comprop.oii.ox.ac.uk/wp-content/uploads/sites/93/2019/01/Manipulating-Twitter-Traffic.pdf [Accessed 20 October 2019].

Pennycook, G., & Rand, D. G. (2018). Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. *Journal of personality*.

Shao, C., Ciampaglia, G.L., Varol, O., Yang, K., Flammini, A. & Menczer, F. (2018). The spread of low-credibility content by social bots. *Nature Communications* 9(1), 1–9.

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1), 22-36.

Starkbaum, J., & Felt, U. (2019). Negotiating the reuse of health-data: Research, Big Data, and the European General Data Protection Regulation. *Big Data & Society*, 6(2), 1–12.

Syed, N. (2017). Real Talk About Fake News: Towards a Better Theory for Platform Governance. *Yale LJF*, 127-337. Available at: https://www.yalelawjournal.org/forum/real-talk-about-fake-news [accessed 2 May 2020]

# 6.3 Case Studies on New Technologies' Implications for Democratic Principles

Vesa Vuolle, Emmi Nahi, Aino Hiltunen, Tom Henriksson, Juuli Hakulinen
Faculty of Social Sciences, University of Helsinki

## Abstract

This paper examines how societies could benefit from emerging areas of societally-applied artificial intelligence (AI) such as big data (BD), machine learning (ML), and algorithmic governance (AG). Furthermore, the paper outlines the challenges these technologies pose to key principles of democracy such as legitimacy, privacy, democratic governance and freedom. The research question is answered by analysing five topical cases: the emerging social credit system in China, the experiment with AI in social services conducted by the City of Espoo, the use of Big Data and AI in UN-led humanitarian efforts, the use of algorithmic targeting in elections, and the use of AI in state surveillance. We conclude that while the aforementioned technologies will likely provide societies with many opportunities, they may also pose serious threats to democratic principles. We recommend further research on the social implications of AI, BD, ML and AG so that adequate regulative action can be prepared to address these threats.

*Keywords*: Big data, algorithmic governance, artificial intelligence, AI, machine learning, democracy, social credit score, China, micro-targeting, digital surveillance

This paper examines how societies could benefit from emerging areas of societally-applied artificial intelligence (AI) such as big data (BD), machine learning (ML), and algorithmic governance (AG). Furthermore, the paper outlines the challenges these technologies pose to key principles of democracy such as legitimacy, privacy, democratic governance and freedom. AI technology is rapidly proliferating around the world. What makes it difficult to examine is that it is not one specific technology, but more of an integrated system that incorporates information acquisition objectives, logical reasoning principles and self-correcting capacities (Feldstein 2019a). In our text, we use a very basic definition of AI as the simulation of human intelligence processes by machines which include learning, reasoning and self-correction (e.g. Poole and Goebel 1998; Russell and Norvig 2003). There is a categorized division between weak and strong AI, in which a weak AI system is designed for a particular task and a strong AI system exhibits generalized human cognitive abilities. In this paper we focus on weak AI. By machine learning we mean a subset of AI. Furthermore, we define Big Data as a field that explores ways to analyse data sets that are too large or complex to manage using traditional data-processing software. From a societal point of view, legitimacy is a central principle: what is considered to be a justified use of AI, machine learning and big data? How can individuals' rights be guaranteed?

According to Habermas (2001), the modern conception of democracy and its relation to law consists "of norms that are produced by a lawgiver, are sanctioned by the state, and are meant to guarantee individual liberties"; moreover, citizens' democratic self-determination can only be realised through structural properties and media such as law, which ensure liberty (Habermas 2001, 766). Five cases will be presented to analyze the problem. First, we will explore the use of big data and algorithmic governance in the context of the social credit system in China. In the next section, we cover how machine learning is utilized in social services, and contemplate the question of structural discrimination and implicit bias in algorithmic governance. In the third section, we will interrogate how big data is used in humanitarian efforts by the United Nations. The fourth section focuses on the case of Brexit, and how big data was used to influence the referendum, paying particular attention to how the case illustrates recurring patterns of influencing the public with algorithmic microtargeting in modern elections and what implications this has on the future of deliberative democracy. The final case looks at the use of AI in governance surveillance, examining the case of Palantir. We conclude by stating that while the aforementioned technologies will likely provide societies with many opportunities, they may also pose serious threats to democratic principles. We recommend further research on the social implications of AI, BD, ML and AG so that adequate regulative action can be prepared to address these threats.

## China's social credit system – Juuli Hakulinen

This section examines the social credit system in China. China is not a democracy, so the focus here is on how new technologies are used to tighten the grip of the regime.

Does the system work? Are new technologies only good for the government or is there something to gain for the people? This section aims to serve as a case study of how big data has arguably been the most useful for authoritarian regimes and is used in the least restricted way. It seeks to demonstrate the extent to which technologies can be used to control and police citizens and analyse this practice through the lens of legitimacy, or to what amount the system is acceptable for Chinese citizens.

China has wanted to create a Social Credit Score (SCS) since the 1940s but was unable to do so until technology had advanced sufficiently (Ramadan 2018, 93). Before the development of new technologies, systems called *Dang'an* and *Hoku* provided ways to monitor citizens. Now CCTV and Internet censorship is used in big data analytics. China established a Cyberspace Administration in 2014, in the framework of which 500 smart cities cooperate with IT firms to better their services, CCTV cameras are abundant (Sky Net Project) and facial recognition technology is at a point where an individual can be followed almost anywhere (Fang, et al 2018, 419-420).

What is meant by SCS here is a "rating for the consumer derived from his/her 'position in a social structure based on esteem that is bestowed by others' [Hu and Van den Bulte 2014, 510]" (Quoted in Ramadan 2018, 93). The process is still under way and no cohesive all-encompassing SCS has been created. Instead, the system now in place is more of an ecosystem that is comprised of government agencies and commercial firms (Creemers 2018; Liang, Fan, Vishnupriya Das, Nadiya Kostyuk, and Muzammil M. Hussain 2018).

Liang et al. studied China's Social Credit Scoring in their article (Liang et al. 2018) and concluded that it perhaps focuses more on financial and commercial activities than political ones. A large portion of data is gathered by commercial firms, which operate in the field of big data. The market size is estimated to be approximately $2.5 billion in 2016 and is expected to grow (ibid. 415). Despite the prevalence of scoring in China being about financial affairs, Backer and Catá (2018) argue that scoring is inseparable from the state. The state uses these techniques to spy on citizens and make them act in accordance with what is ideologically acceptable.

While SCS is the state's ideological tool, it is also a technique that is inspired by traditional commercial credit rating, which is applied to social behaviour. Traditional credit scoring systems are numerical systems, in which the decision to classify someone's credit worthiness is predicted, ruling out personal judgement (Abdou and Pointon 2011). In China, the aim is to use the available data in order to create a credit scoring for each of the 1.4 billion individuals in the country. This score will determine whether an individual can be granted access to certain benefits or if they will be punished. Government affairs, judicial affairs, social activities, and commercial behaviour are spied on with the help of data gathered in massive amounts (Fang et al. 2018, 416).

Fang et al. (2018) claim that in the literature on information and communication technologies (ICTs) discussing authoritarian regimes, the approach has been either optimistic or pessimistic. Some scholars see the potential of ICTs to empower citizens, and others as a way for states to exert more power over their subjects.  These two camps can also be found in literature about China (ibid. 416). In a way, technology can help citizens free themselves from the grip of the state. This has happened to some extent in China. Memes or emojis can be used to circumvent some of the state's censorship. It has also been argued, however, that the ability of Chinese citizens to use new technologies to "fight back" has decreased (ibid. 434).

We can ask whether China's SCS is any different from western data gathering and more conventional credit rating. The big difference between the West and China in this matter is that China aims to collect all data, on everybody, all the time (Fang et al. 2018, 417). This means that every piece of data can be linked to an individual. This differs from samples that are collected about a population as opposed to about individual citizens. Another difference is that this new way of collecting data is not visible, unlike data collection practices before. In China, as opposed to the West, the state and private actors work together to gather information. Platforms are mostly state owned and not private enterprises (ibid. 429).

Maybe the most obvious problem from the point of view of citizens' rights, despite the fact that there is no democracy in China, is the loss of privacy. Especially personal privacy is an issue, because private data is readily available about individual people. Also, for complex reasons the people have not been able to use the same techniques for counter-surveillance. There is no civil society to keep the state in check (Fang et al 2018, 420). For now, however, SCS is mostly a tool for the government to blacklist individuals. It is not yet a sophisticated way of monitoring and anticipating the actions of individuals. What remains to be seen is whether China will seek to export its system to other democratic countries. At least one instance of such behaviour has been observed, when a Chinese state representative attempted to get Canada to accept their Sesame credit rating (ibid. 435).

It is unclear whether China's SCS will backfire (Fang et al. 2018, 435). What is better documented is the widespread acceptance the system seems to have among well-off citizens (Kostka 2019). This can be explained by the fact that these people generally benefit from the system and already respect the rules of the government and the party. In that position an individual might not care so much about the loss of their privacy.

Based on the information available, the most obvious problems with China's SCS for the rights of citizens seem to be the loss of privacy and the illegitimacy and lack of transparency in social credit scoring. No one knows exactly what will be monitored and how. This could worsen into a panopticon-like situation where people start to monitor themselves. The way in which the state will develop the system is also unclear. Still, the population largely accepts the system, as it offers benefits for model citizens. In a competitive society it is understandable that some might not want to defy a system that gives them a head start at life. Like in other cases examined in this paper, it seems that the Chinese system, which is still taking shape, needs to balance between what is useful and efficient for the government and the party, and what people are willing to accept. In other words, the legitimacy of the SCS depends on how it serves citizens' interests.

## Machine learning, social services, and structural discrimination – contemplating the question of implicit bias in algorithmic governance – Vesa Vuolle

One of the key roles of a social worker is to assess customers' needs for social services. While the theories and practice are familiar with many risk factors in social services, social workers have to rely on subjective reasoning to assess them. However, in ever more areas of life, algorithms are being introduced to substitute the judgments exercised by identifiable human beings who can be held accountable (Crawford 2019, 1). Widespread application of artificial intelligence in healthcare has been anticipated for half a century (Hinton 2018, 1) and now predictive modelling with electronic health record data is expected to drive personalized medicine and improve healthcare quality (Rajkomar et al. 2018, 1). Yet, intelligent though they may be, these algorithms exhibit some of the same biases that permeate society (Howard & Bornestein 2017, 1521). This paper contemplates the possibility, that while artificial intelligence can be of great help in one domain, it could simultaneously reproduce and strengthen regressive prejudices, such as racism and sexism.

The city of Espoo and the Finnish ICT-company Tieto Corporation have developed a social services experiment, in which artificial intelligence can pick service paths out of an enormous mass of service data by grouping risk factors that trigger the need for intensive and expensive social services, if found in the same person. Those backing the project observe that, "The Espoo experiment proves that artificial intelligence recognises those who need support" (Espoo 2018). The algorithm is unique because it utilizes public data, as opposed to data gathered by private companies. The experiment, which started in 2017, has since helped social workers to detect the likelihood of a person becoming the subject of social services. The algorithm found approximately 280 predictive risk attributes (Valtioneuvoston selonteko 7/2018, 2019). These cases, wherein public administration policy-makers utilize big data, have reshaped how social services can be viewed, although not without controversy (Dunleavy 2016). According to Williamson (2016, 136), the basic functional principles of emerging big data technologies can be seen to be starting to structure public policy guidelines. This resembles liberal proceduralism, but with an important difference: the rationale behind policies initiates from Black box type semi-autonomous machines, as opposed to politicians or civil servants. This is a fundamental problem posed by mechanised decision-making, as it touches on the basis of political legitimacy in any liberal regime (Crawford 2019, 1). Further, in this respect, algorithms exercise social power that can influence patterns of human agency (Neyland & Möllers 2017).

Attributes associated with high-cost social services patients may include behavioural health problems or socioeconomic factors such as poverty or racial minority status (Rajkomar et.al. 2018, 3). This is the sort of information that Pasquale

(2015) describes as information that determines our standing in the reputational economy. If an algorithm learns to screen people based on these factors, doesn't it then reinforce the structural powers that currently patronize and oppress many minorities, making it more difficult for those born with inferior conditions to beat the odds of having limited futures? It would seem that this kind of analytical property (or bias) of an algorithm strengthens inequality by categorizing certain sociocultural features as adverse and forecasting future use of social services. In an example, Angwin et.al. (2016, 32) attest that some of the relevant computing systems used by the US justice system unfairly discriminate against African Americans by at times ranking them as being more dangerous than their white counterparts when the reverse may be true. This is in part due to the way a machine learns; it merely predicts future outcomes based on historical data which it was trained on.

According to Borenstein and Howard (2017, 1524), logic and evidence should be transparently presented when seeking to make the case that treating an individual or group differently is appropriate. However, according to Crawford (2019, 3), the decisions made by an algorithm are often not explainable, even by those who wrote the algorithm. The algorithm is free of subjective considerations (or deviations) a social worker might take into account. As these algorithms evolve into advanced artificially-intelligent agents, intertwining with our physical world, the negative ramifications of bias only increase (Borenstein & Howard 2017, 1522). Therefore, with the inauguration of the technology-driven preventive system, all subjective and humane contemplations are being reduced. A person who is born with an intersectional minority status would most likely be screened as a high costing patient of social services by an algorithm, even if that individual had an extraordinarily high IQ, energetic drive, loving family and the will to make the best of these.

Consider what Carpenter and Guarino, Borenstein and Howard (2017, 1524) acknowledge, that allegations of racism and sexism have permeated the conversation about AI as stories surface about search engines delivering job postings for well-paying technical jobs to men and not women, or providing arrest mugshots when keywords such as "black teenagers" are entered. Ali (2019) has even written about algorithmic racism and algorithms as tools for upholding white hegemony. This is not, however, the first time a machine goes rogue. Many remember the Microsoft chatbot Tay, who was taught by Twitter users to swear, make racist remarks and inflammatory political statements (Wakefield 2016).

In these respects, a Foucauldian framework of biopower, ascribing social power through discourse and surveillance would provide useful analysis to the theme (Foucault 2007). Lloyd (2019, 1479) suggests that the creation and circulation of algorithms produces a discourse of truth that may not be refuted because the thinking behind them is not made available. This raises questions of trust, legitimacy, bias, and credibility towards the semi-autonomous machines in health governance.

It would be fascinating to survey social workers in Espoo, and whether they consider that algorithms have discovered new customers previously unknown to social workers. This could include for instance noticing poverty in an otherwise affluent neighbourhood or pre-emptive interventions to entrepreneurs whose business is on its way to going bankrupt. Alternatively, the algorithm could reverse decisions of service paths social workers have already suggested, by detecting implicit features that have protective qualities, such as soon-to-graduate students who are poor in resources but rich in assets.

The extent to which algorithms come to impact the future of social services remains to be seen. The city of Espoo has been struggling with the European Union's General Data Protection Regulation but is determined to continue the use of its one-of-a-kind algorithm (Valtioneuvoston selonteko 7/2018, 2019). The issue of algorithms as part of social services introduces many moral considerations, of which implicit prejudices related to the colour of skin, gender or a postal code in a poor neighbourhood are not the least. At the end of the day, it is the unique and critical interplay between nature and nurture that makes us who we are. No one should be doomed because of these, nor should anyone be left outside safety nets just because their historical data does not predict future deprivation.

## Big Data for UN-led humanitarian efforts – Aino Hiltunen

This section examines how big data is used in humanitarian efforts by the United Nations. The section focuses on UN-led efforts to harness big data for humanitarian endeavors through the UN Global Pulse initiative, and identifies the challenges and pitfalls that the UN-backed approaches may encounter.

The United Nations Global Pulse is an innovation initiative of the United Nations Secretary-General on big data aiming to harness big data "safely and responsibly as a public good" (UN Global Pulse 2018a). The mission is to "accelerate discovery, development and adoption of big data innovation for sustainable development and humanitarian action" (UN Global Pulse 2018b). The initiative works through "innovation labs", consisting of data scientists, engineers, designers, social scientists, communication experts, and data privacy as well as legal experts, who work with humanitarian and development actors to design and implement innovation programs, share knowledge and produce reports, technical papers and project briefs (UN Global Pulse 2018a, 10).

Existing literature on the use of big data for development purposes focuses especially on applications of big data for mitigating crisis situations. Ali et al. (2016) highlight the reactive application of big data tools with an example on how the Syrian refugee crisis was mitigated by UNHRC and non-profit volunteer organizations by creating an online interactive map, through which real-time information was made available to refugees about aid agencies, organizations, their functions and operating times as well as capacities (Ali et al. 2016, 10). Ali et al. state that the "real promise" of big data for development purposes might lie in predictive analytics, where humanitarian emergencies could be avoided or mitigated before they actually happen (Ali et al. 2016, 10), in contrast to reactive purposes. The authors do not explore the idea further besides referring to businesses that use previous data in predicting customer behavior.

In a similar vein,  Karlsrud (2014) argues that big data can be utilized to provide insight to crisis and disaster response in order to react effectively to quickly transforming situations. This is also echoed in the research by Qadir et al. (2016) who examine how big data analytics can be utilized during emergencies to mitigate crisis situations, or even help to avoid a crisis altogether. Qadir et al. identify eight fields of humanitarian aid and development where data-driven responses are transforming crisis response practices: epidemic crises, natural disasters, crowd control issues, terrorist attacks, civil wars, public violence and other disaster situations, such as infrastructural failures and industrial accidents (Qadir et al. 2016, 2).

The existing literature seems to focus especially on crisis mitigation through the means of big data technology. The research focus has been on the positive, reactive and preventive possibilities that big data has to offer, and rarely discusses critically the use of big data in development and humanitarian settings in detail. The existing literature does not assess or analyse the United Nations initiative, but merely mentions it as an actor in the field for utilizing big data for development purposes and humanitarian action. Therefore, research should be conducted on the UN initiative itself in order to gain broader insight on its successes as well as pitfalls.

Using big data for development and humanitarian purposes within the UN is not without challenges. Privacy is one of the biggest concerns for the usage of big data in general, and the field of development and humanitarian action by the United Nations is no exception. The question of privacy poses one of the most pressing ethical challenges: with large amounts of data collected on individuals, it should be a top concern that data should not be abused for personal or financial gains (Ali et al. 2016, 21). The UN Global Pulse has cooperated with corporations, such as Orange, in an attempt to engage the private sector in "data philanthropy" (Kirkpatrick and Vacarelu 2018). The aim of the UN Global Pulse is to work with the private sector in order to put the privately collected data to use in order to promote sustainable development goals. What seems to be problematic, however, is that there is no standardized regulation for sharing data for the "common good". In addition, the projects should be critically assessed to ensure that the data is not used for financial gain under the pretense of humanitarian action and development. The legitimacy of using private companies to gather data that is then used to promote UN goals can also be questioned. Does the reliance on privately gathered data create possible dependencies on how the data is then utilized? How do we ensure that the data is not used to promote the corporate agendas in the shadow of the rhetoric of development and humanitarian action? How do we make sure that the data is collected in a transparent fashion, considering the ethical challenges that the age of big data might pose? The UN framework and the social development goals are not without issues themselves, for example when related to neoliberal governance that promotes market-oriented solutions to development and humanitarian questions.

Qadir et al. highlight that even though big data can be useful in providing insight to development issues and crisis situations in terms of a humanitarian response, it is not the only standing solution to the complex issues faced (Qadir et al.

2016, 19). Ali et al. identify the challenges to the use and implementation of big data for development purposes from technical and ethical perspectives (Ali et al. 2016, 19). The technical challenges include bias and polarization emergent from the personalized content predicted by past behaviour and double responses from crowdsourcing, wherein for example aid agencies take on the same problem at the same time without coordination and complementary action.  This could be a problem within the UN Global Pulse context, if a failure to coordinate with other actors would materialize. Laura Mann (2018) uses a political economy perspective to the use of big data for development, and concludes that data is often extracted from the Global South, especially African contexts, for the use of "humanitarian purposes". According to Mann, the approach shows how data is extracted by multinational corporations, which then aim to become "data custodians" of emerging economies in the global south. This approach incorporates global power relations to the analysis of big data for development and humanitarian action, which should not be disregarded. When data is extracted by companies, and emerging economies become increasingly digital, data becomes a source of power and economic governance, which has manifold implications. Big data representativeness can also cause challenges, since equal access to information technology, mobile phones and the Internet especially in contexts of crisis are not certain (Hilbert 2016, 149). Using big data for development purposes requires  nuanced understanding, analysis and awareness on how technologies might have the capacity to enhance or impede capabilities (Hilbert 2016, 140). When the UN is utilizing big data, it needs to be considered how the use of big data might lead to increasing state and corporate control, and how ethical concerns are assessed and incorporated into policy and utilization frameworks.

This section of the research paper examined the role that big data play in UN-led efforts in sustainable development and humanitarian action, looking at the existing challenges and opportunities. Concluding the discussion, big data provides numerous possibilities and ways forward in the context of the UN action, but it is not without challenges.

## Brexit and microtargeting: a threat or an opportunity for democracy? – Tom Henriksson

The 2019 film *Brexit: The Uncivil Wa*r addressed the emerging issue of microtargeting in elections and referenda. In a key scene of the movie, Dominic Cummings meets the representative of a data mining company. The representative claims that his company can algorithmically target ads to voters. Furthermore, he claims that ads can mobilize 3 million people to vote.

In this paper, "microtargeting" refers to targeting online ads to people based on an analysis of their preferences and behaviour by artificial intelligence. Whether microtargeting was decisive for the result of the Brexit referendum would be an interesting topic, but it is clearly outside of the scope of this essay. Instead, this section will discuss the implications of microtargeting to democratic principles such as information, privacy, transparency, public deliberation and equal chances of success in elections. I will present the arguments of Borgesius et al., as their approach is the most systematic available, and also introduce readers to the research of Heawood and Benkler. Brexit is the case used here to elaborate these authors' research findings.

Borgesius et al. divide their analysis of the pros and cons to citizens, parties and the public opinion (Borgesius et al. 2018, 84). For citizens, the benefits include increased participation and more informed choices, since targeted ads can connect people with agenda points they care about. Ads on social media can also reach audiences that do not follow mass media (Borgesius et al. 2018, 84-85). Political parties, on the other hand, can run cheaper and more efficient campaigns. Online microtargeting is sometimes cheaper than traditional methods such as newspaper ads (Borgesius et al. 2018, 85-86). Borgesius et al. argue that voters easily become overloaded with information in the "marketplace of ideas" during elections, but microtargeting can work as a sieve that filters the information most relevant to them. Thus, the public could be more informed (Borgesius et al. 2018, 86).

On the other hand, citizens face a threat to their privacy, as online microtargeting involves collecting massive amounts of sensitive data that is vulnerable to breaches and liable to misuse. Microtargeting could be used to manipulate voters all while increasing polarization and spreading lies. Instead of increasing participation, the turnout of undesired voters could be suppressed. Some groups could be ignored if their interventions are, for example, more oriented to provoking chaos or than furthering political aims (Borgesius et al. 2018, 87-88). For parties, professional microtargeting may end up costly, making the competition harder for emerging challengers. Intermediaries such as data companies could become increasingly powerful (Borgesius et al. 2018, 88-89). Political parties could also present themselves differently to each voting segment, without revealing a wider programme or the real priorities of the party. As interest in the overarching issues decreases, events of public deliberation such as debates will have a smaller following (Borgesius et al. 2018, 89).

Jonathan Heawood raises similar concerns for democracy such as concerns about privacy, incompatible promises to different segments, as well as foreign influence (Heawood 2018, 431-432). However, his key contribution is to highlight a specific feature of microtargeted ads: unlike e.g. newspaper ads, they cannot be reached or seen equally by everyone: "microtargeted political adverts are, in this respect, a 'pseudo-public' form of discourse" (Heawood 2018, 430). Microtargeting thus happens outside the public sphere as conceived in the Habermasian sense. In a scene of *Brexit: The Uncivil Wa*r, the Leave.eu campaign is microtargeting voters on social media with objectively false claims. However, journalists are not reporting anything, because they are not seeing the ads. As Heawood puts it, "claims made in private cannot be corrected in the marketplace of ideas" (Heawood 2018, 431).

However, Heawood warns against naïveté. Criticizing Borgesius et al., he believes that "this implicit distinction between a 'good' old public sphere and a 'bad' new digital sphere is a bit of a fantasy" (Heawood 2018, 432). Any newspaper that is not free conceals some key political information from the public. Even public broadcasters are run by people: the BBC's tendency for "balance" is ridiculed in *Brexit: The Uncivil War*, when the BBC is presented as giving equal time to analyses by Nobel prize winners and outright liars. According to Heawood, we should be equally focused on solving the old problems of the analogue public sphere as we are on addressing the new challenges of the digital public sphere: ideally, the emergence of the latter may make the problems we already had with the former more visible (Heawood 2018, 433-434). Organising both traditional and social media townhall meetings might be one way to enliven both spheres.

As a way to counter the potential threats of microtargeting, Borgesius et al. propose more research on microtargeting. Possible regulation could include public repositories of each ad, limits for campaign budgets or even a blanket ban on microtargeting (Borgesius et al. 2018, 92-95). While Benkler et al. (Benkler, Faris, & Roberts 2018) have a significantly more positive view on microtargeting than Borgesius et al. and Heawood, they recommend that "individually tailored, or too narrowly targeted advertising techniques … be constrained in the political context" as they undermine the perception of legitimacy of elections (Benkler et al. 2018, 279). Benkler also supports the idea of a public ad repository (Benkler et al. 2018). For now, intermediaries such as Twitter and Facebook can be considered to have both social responsibility and power outside of democratic processes. Twitter recently banned all political ads from its platform, while Facebook is introducing stricter rules (Hunnicutt 2019). A blanket ban may be problematic from the perspective of free speech; the sincerity of Facebook can be questioned, as ads remain a key source for social media platforms.

Finally, I will add some of my own thoughts. First, I would like to propose a reverse notion compared to Borgesius et al. They note that in continental Europe, the threats of microtargeting are less relevant than in the US or the UK. Legal protections such as the European data protection law don't exist in the US, and both the US and the UK systems feature majoritarian processes (elections in which only one candidate wins in a constituency) (Borgesius et al. 2018, 89-91). If majoritarian systems are more vulnerable to voter manipulation, maybe this should be added to the list of pros and cons in the discussion of which system is better. Second, I would like to note that the Brexit referendum was somewhat unique, inasmuch it presented a simple dichotomy. Instead of introducing referenda with oversimplified choices—think about David Cameron promising a referendum on the simple issue of the membership of the European Union—political leaders should provide the public with a more incremental process. If simple politics is vulnerable to manipulation, let us not oversimplify politics.

Third, Mancosu et al. have demonstrated that older votes were more likely than young voters to respond to Facebook status updates designed to mobilise them during the Brexit referendum (Mancosu & Bobba 2019). Could it be that young people, as social media natives, are less vulnerable to being targeted on social media, with a more matter-of-fact relationship to ads? While microtargeting will likely become more sophisticated in the following decade, social media will eventually become less of a novelty. As younger generations get older and eventually become increasingly socialized in the political system, could the effects of microtargeting lose significance?

## Artificial Intelligence and government surveillance – the case of Palantir – Emmi Nahi

Awareness of cyber vulnerabilities is in its infancy (Bernett 2019). It has, however, become an increasingly interesting topic. AI's impact extends to transforming patterns of governance, not only by giving governments new capabilities to disrupt elections and elevate false information but also to monitor their citizens and shape their choices (Feldstein 2019, 5).

In this case study we examine the role of AI in government surveillance. This requires a brief overview on international human rights law and how it applies in the current digital environment, particularly in light of the increase and changes in surveillance technologies. Privacy is a fundamental principle and recognized under international human rights law (OHCHR 2018, A/HRC/27/37). It is essential to human dignity and reinforces other rights, such as freedom of expression and information (Necessary & Proportionate 2014). Because surveillance may only be justified when it is prescribed by law and necessary to achieve a legitimate aim, proportionality is the key principle. States must justify every intrusion in an individuals' private life; it needs to happen for a legitimate reason, and the surveillance measure cannot be more intruding than necessary to meet the needs.

A growing number of states are deploying advanced AI surveillance tools to monitor, track and surveil their citizens to accomplish a range of policy objectives, some violating human rights, some not and—what is more important—most of them falling somewhere in the middle (Feldstein 2019). According to the AI Global Surveillance (AIGS) Index by the Carnegie Endowment for International Peace, at least 75 countries out of the examined 176 are actively using AI-technologies for surveillance purposes, liberal democracies being the major users. The index does not distinguish between legitimate and unlawful uses of AI. The findings include for example smart city platforms, facial recognition systems and smart policing. China is the major driver of AI surveillance with its companies such as Huawei, Hikvision and ZTE. However, the US does not fall short. AI surveillance technology supplied in US companies is present in 32 countries. The most significant firms are IBM, Cisco and the focus of this case study, Palantir.

Palantir is a powerful but not well-known CIA-backed data gathering and analysing start-up owned by Peter Thiel - a billionaire Pay-Pal co-founder, Facebook investor, a latter-day Trump ally, and presidential adviser (Biddle 2017). Palantir has worked years to boost the global dragnet of the National Security Agency (NSA) and CIA as well as elements of the US military; and it was in fact co-created together with American spies (Biddle 2017, 2; Posner 2019). Although Palantir does not mask its ambitions, which is to sell its services to the US government, it does refuse to name its governmental clientele, despite the fact that it has landed around $1,2 billion in federal contracts since 2009 (Biddle 2017), consequently making its operating extremely non-transparent.

What Palantir has been the most criticized for is its technical support for US immigration enforcement practices (Posner 2019). It has provided services for the U.S. Immigration and Customs Enforcement (ICE) which has caused protest marches in several cities and frustration among its own employees (e.g. Bort 2019, Woodman 2017, Carroll 2019).

Palantir has been providing ICE with software since 2014, but denied that its technology is used by the part of ICE that handles family separations and deportations (Bort 2019). Conflicting information was reported by the Intercept (Woodman 2017) claiming that Palantir has created an intelligence system called Investigative Case Management (ICM) which is deployed by ICE and assists in President Trump's efforts to deport immigrants from the US. According to government funding records (Department of Homeland Security record 2014), ICE awarded Palantir a $41 million contract to build and maintain ICM, and identifies the program as "mission critical to ICE". ICM is accessible by both ICE's Enforcement and Removal Operations (ERO)—the federal government's primary deportation force—and Homeland Security Investigations (HSI), despite Palantir's claims otherwise (Woodman 2017).

One of these services is the FALCON system, "a database and analytical platform built by Palantir that HSI agents can use to track immigrants and crunch data on forms of cross-border criminal activity" (Woodman 2019). Falcon costs taxpayers $39 million (USA Spending). In a nutshell, ICE uses Palantir's software to create "digital dragnets" of individual people in an attempt to predict crime before it happens (Haskins, 2019). The data includes emails, phone numbers, addresses, social security numbers, business relationships, travel histories captured by license plate cameras, and social networks, among other types of information (e.g. Haskins 2019, Waldman et al. 2018). The data presents the connections in colourful, easy-to-interpret graphics that are very much appreciated by US spies and special forces and helped planners avoid roadside bombs, track insurgents and even hunt down Osama bin Laden (Waldman et al. 2018). The military success, justifiable or not, led to federal contracts on the civilian side. Now Palantir sells its services to make a powerful surveillance system at NSA even more powerful, bringing clarity and slick visuals to surveillance data (Biddle 2017, 6).

Moreover, after the privacy hustle caused by Edward Snowden, Palantir quickly denied its connections to the NSA spy program called PRISM which, however, shared an unfortunate code name with one of its own software products (Biddle 2017, 4). Palantir's website includes a whole section of "Privacy & Civil Liberties" but does not get to practicalities.

Well-established legal principles in monitoring communications that were created before the public adoption of the Internet have decreased in recent decades and the application of legal principles in the new technological context has become unclear (Necessary and Proportionate 2014).  The principle of legitimacy is closely related to transparency. As a privately held company, Palantir is not required to reveal much about its finances or operations (Posner 2019). Instead, it has profited from governmental "data-analytics"— or mass-surveillance—in the name of predicting crime or terrorism, but in fact endangering human rights. Palantir has access to classified military data, facial recognition cameras across the country and at the borders, as well as the complete trust and cooperation of the federal government and hundreds of local law enforcement agencies (Greene 2019). According to immigration activist group Mijente, it is "a surveillance machine capable of tracking anyone and everyone".

The state's struggle to balance security interests with citizens' liberties is the main topic of this section. Strong security and the use of AI surveillance are closely related: out of the top 50 military spending countries 40 deploy AI surveillance technology (Feldstein 2019b, 11). It is unclear, to what extent surveillance deployments are covered in national or international law, let alone whether the actions meet the necessity and proportionality standard. This is increasingly interesting as the major users of AI surveillance are liberal democracies, where the demand for legitimacy is high, regardless of the purpose of the surveillance. The explosion of digital communications content and the falling cost of storing and mining large sets of data make surveillance by states possible at an unprecedented scale. Palantir's high profile and often controversial business activities provide an instructive case study on the issue of state surveillance and consequently the Trump Administration's practice of data analysis facilitating mistreatment of asylum seekers and other migrants.

## Conclusion

This paper claims that while the new smart technologies will likely provide societies with many new opportunities, they may also pose serious threats to democratic principles and social cohesion.

The first case dissected China's SCS program, showing that it engenders citizens' privacy, while undermining the democratic ideals of legitimacy and transparency. As with other cases examined in this paper, it seems that the Chinese system needs to balance between what forms of monitoring people are useful and efficient for the government and the party, and what people are willing to accept. This raises the concern of legitimacy. The theme of transparency continued in the second case. By examining the literature related to the ethics of AG, the review suggested that the black box type algorithms used in artificial intelligence and machine learning should be coded in a socially and ethically progressive way. As algorithmic governance gains more foothold in the future, resulting policy suggestions need to be able to see through, prevent and reverse any structural discrimination humans have practiced in the past.

Following the first two nation-state oriented cases, the paper took a more global view by looking at how The United Nations aims to use big data for development and humanitarian purposes through the UN Global Pulse initiative. Here again, the issue of privacy posed one of the most fundamental questions regarding the use of big data for development and humanitarian purposes, since the data used is most often collected by private companies. It became evident that problems will arise since there is no existing framework on how to regulate data sharing for "the common good", or to ensure that data used for humanitarian purposes is not used for financial gain. The UN's focus on big data for development should, therefore, assess whether the data used for humanitarian purposes is collected in a transparent fashion, and provide tools for the analysis and assessment of safe and responsible data use.

The fourth case discussed the emerging phenomenon of microtargeting voters in elections and referenda. Microtargeting is an example of how new technology may both invigorate and challenge democratic principles. Targeted information may help the public find the information most relevant to them in the "marketplace of ideas", helping more people to get involved in the political process. However, microtargeting also facilitates manipulation, voter suppression and the spread of misinformation. The key to addressing the challenge is improving our understanding of the meaning of the public sphere, both digital and analogue. Regulation should safeguard transparency, privacy and access to verifiable information. Oversimplification of complex political issues should be avoided.

Finally, the fifth case explored how international and national law has not kept up with technological development, causing new technological activities to fall in a juridically grey area. Further, the U.S. government has a variety of tools at its disposal to surveil its citizens, many of which the public is unaware of. An epitome of this is the case of NSA and Palantir, a privately-owned company. This paper has shown that AG, ML, and BD can be used to amplify the effectiveness of policies, whether they are democratic or authoritarian, benevolent or atrocious. The digital revolution is likely to change how social relations and wider societal contexts are perceived in academia, administrations and elsewhere. We, therefore, recommend further research on the social implications of AI, BD, ML, and AG so that adequate regulative action can be prepared to address these threats.

## References

Abdou, Hussein A., and John Pointon. 2011. "Credit Scoring, Statistical Techniques and Evaluation Criteria: A Review of the Literature." *Intelligent Systems in Accounting, Finance and Management* 18(2–3): 59–88.

Ali, Anwaar et al. 2016. "Big Data for Development: Applications and Techniques." *Big Data Analytics* 1(2): 1–24.

Ali, Syed Mustafa. 2019. "'White Crisis' and/as 'Existential Risk,' or the Entangled Apocalypticism of Artificial Intelligence." *Zygon*® 54(1): 207–24.

Backer, Larry Catá. 2018. "Next Generation Law: Data-Driven Governance and Accountability-Based Regulatory Systems in the West, and Social Credit Regimes in China." *Southern California Interdisciplinary Law Journal* 28(1): 123–72.

Benkler, Yochai. 2018. *Network Propaganda : Manipulation, Disinformation, and Radicalization in American Politics*. Oxford University Press.

Bernett, Jackson. 2019. AI is breathing new life into the intelligence community. *Scoop News Group,* 21 Aug. https://www.fedscoop.com/artificial-intelligence-in-the-spying/ (October 28, 2019).

Biddle, Sam. 2017. "How Peter Thiel's Palantir Helped the NSA Spy on the Whole World." *The Intercept*. https://theintercept.com/2017/02/22/how-peter-thiels-palantir-helped-the-nsa-spy-on-the-whole-world/ (November 4, 2019).

Borgesius, Frederik J. Zuiderveen et al. 2018. "Online Political Microtargeting: Promises and Threats for Democracy." *Utrecht Law Review* 14(1): 82–96.

Bort, Julie. 2019. "Palantir's Tech Was Used by ICE in the Controversial Arrests of 680 People at a Mississippi Chicken Farm According to an Immigrants' Rights Group." *Business Insider*. https://www.businessinsider.com/activist-group-targets-palantir-over-controversial-ice-raid-2019-10 (November 4, 2019).

Carnegie Endowment for International Peace*:* AI Global Surveillance Index. https://carnegieendowment.org/files/AI_Global_Surveillance_Index1.pdf (November 3, 2019).

Carroll, David. 2019. "China Embraces Its Surveillance State. The US Pretends It Doesn't Have One." *Quartz*. https://qz.com/1670686/the-us-has-a-lot-in-common-with-chinas-surveillance-state/ (November 4, 2019).

Crawford, Matthew B. 2019. "Algorithmic Governance and Political Legitimacy." *American Affairs* 3(2): 73–94.

Creemers, Rogier. 2018. *China's Social Credit System: An Evolving Practice of Control*. Rochester, NY: Social Science Research Network. SSRN Scholarly Paper. https://papers.ssrn.com/abstract=3175792 (September 26, 2019).

Department of Homeland Security report: ICE Investigative Case Management System. https://www.fbo.gov/index?s=opportunity&mode=form&id=36fb3b697a2ccb4ec7084b4e0ec6cdb9&tab=core&_cview=1 (November 4, 2019).

Dunleavy, Patrick. 2016. "Big Data' and Policy Learning." In *Evidence-Based Policy Making in the Social Sciences: Methods That Matter*, eds. Gerry Stoker and Mark Evans. Bristol University Press, 143–68. http://www.jstor.org/stable/10.2307/j.ctt1t89d4k (October 3, 2019).

"Espoo Experiment Proves That Artificial Intelligence Recognises Those Who Need Support." 2018. *espoo.fi*. http://www.espoo.fi/en-US/City_of_Espoo/Innovative_Espoo/Espoo_experiment_proves_that_artificial_(142925) (October 18, 2019).

"Espoo and Tieto testing artificial intelligence to identify service pathways | Espoon kaupunki - Esbo stad." 2017. https://www.sttinfo.fi/tiedote/espoo-and-tieto-testing-artificial-intelligence-to-identify-service-pathways?publisherId=3385&releaseId=61993408 (October 18, 2019).

Feldstein, Steven. 2019a. "The Road to Digital Unfreedom: How Artificial Intelligence Is Reshaping Repression." *Journal of Democracy* 30(1): 40–52.

Feldstein, Steven. 2019b. "The Global Expansion of AI Surveillance." Carnegie Endowment for International Peace. https://carnegieendowment.org/2019/09/17/global-expansion-of-ai-surveillance-pub-79847 (November 4, 2019).

Foucault, Michel. 2007. *Security, territory, population: lectures at the Collège de France, 1977-78*. New York, Springer.

Greene, Tristan. 2019. "Study: Trump's Paid Peter Thiel's Palantir $1.5B so Far to Build ICE's Mass-Surveillance Network." *The Next Web*. https://thenextweb.com/artificial-intelligence/2019/08/12/study-trumps-paid-peter-thiels-palantir-1-5b-so-far-to-build-ices-mass-surveillance-network/ (November 4, 2019).

Habermas, Jürgen. 2001. "Constitutional Democracy: A Paradoxical Union of Contradictory Principles?" *Political Theory* 29(6): 766–81.

Haskins, Caroline. 2019. "Revealed: This Is Palantir's Top-Secret User Manual for Cops." *Vice*. https://www.vice.com/en_us/article/9kx4z8/revealed-this-is-palantirs-top-secret-user-manual-for-cops (November 4, 2019).

Heawood, Jonathan. 2018. "Pseudo-Public Political Speech: Democratic Implications of the Cambridge Analytica Scandal." *Information Polity: The International Journal of Government & Democracy in the Information Age* 23(4): 429–34.

Hilbert, Martin. 2016. "Big Data for Development: A Review of Promises and Challenges." *Development Policy Review* 34(1): 135–74.

Hinton, Geoffrey. 2018. "Deep Learning—A Technology With the Potential to Transform Health Care." *JAMA* 320(11): 1101.

Howard, Ayanna, and Jason Borenstein. 2018. "The Ugly Truth About Ourselves and Our Robot Creations: The Problem of Bias and Social Inequity." *Science and Engineering Ethics* 24(5): 1521–36.

Hunnicutt, Trevor. 2019. "Twitter Bans Political Ads; Facebook's Zuckerberg Defends Them." *Reuters*. https://www.reuters.com/article/us-twitter-ads-idUSKBN1X92IK (November 4, 2019).

Karlsrud, John. 2014. "Peacekeeping 4.0: Harnessing the Potential of Big Data, Social Media, and Cyber Technologies." In *Cyberspace and International Relations: Theory, Prospects and Challenges*, eds. Jan-Frederik Kremer and Benedikt Müller. Berlin, Heidelberg: Springer Berlin Heidelberg, 141–60. https://doi.org/10.1007/978-3-642-37481-4_9 (October 3, 2019).

Kirkpatrick, Robert, and Felicia Vacarelu. 2018. "A Decade of Leveraging Big Data for Sustainable Development." https://www.un.org/en/un-chronicle/decade-leveraging-big-data-sustainable-development (October 3, 2019).

Kostka, G. (2019). China's social credit systems and public opinion: Explaining high levels of approval. New Media & Society, 21(7), 1565–1593. https://doi.org/10.1177/1461444819826402

Liang, Fan, Vishnupriya Das, Nadiya Kostyuk, and Muzammil M. Hussain. 2018. "Constructing a Data-Driven Society: China's Social Credit System as a State Surveillance Infrastructure." *Policy & Internet* 10(4): 415–53.

Lloyd, Annemaree. 2019. "Chasing Frankenstein's Monster: Information Literacy in the Black Box Society." *Journal of Documentation* 75(6): 1475–85.

Mancosu, Moreno, and Giuliano Bobba. 2019. "Using Deep-Learning Algorithms to Derive Basic Characteristics of Social Media Users: The Brexit Campaign as a Case Study." *PLoS ONE* 14(1): e0211013.

Mann, Laura. 2018. "Left to Other Peoples' Devices? A Political Economy Perspective on the Big Data Revolution in Development." *Development & Change* 49(1): 3–36.

"Necessary and Proportionate," Necessary and Proportionate: International Principles on the Application of Human Rights to Communications Surveillance. 2014.

Office of the UN High Commissioner for Human Rights. 2018. "The Right to Privacy in the Digital Age", A/HRC/27/37

Pasquale, Frank. 2015. *The Black Box Society: The Secret Algorithms That Control Money and Information*. Cambridge: Harvard University Press.

Peretti, Jacques. 2017. "Palantir: The 'Special Ops' Tech Giant That Wields as Much Real-World Power as Google." *The Guardian*. https://www.theguardian.com/world/2017/jul/30/palantir-peter-thiel-cia-data-crime-police (November 4, 2019).

Poole, David L., Alan K. Mackworth, and Randy Goebel. 1998. *Computational Intelligence: A Logical Approach*. New York: Oxford University Press.

Posner, Michael. 2019. "How Palantir Falls Short of Responsible Corporate Conduct." *Forbes*. https://www.forbes.com/sites/michaelposner/2019/09/12/what-companies-can-learn-from-palantir/ (November 4, 2019).

Qadir, Junaid et al. 2016. "Crisis Analytics: Big Data-Driven Crisis Response." *Journal of International Humanitarian Action* 1(1): 12.

Rajkomar, Alvin et al. 2018. "Scalable and Accurate Deep Learning with Electronic Health Records." *npj Digital Medicine* 1(1): 18.

Ramadan, Zahy. 2018. "The Gamification of Trust: The Case of China's 'Social Credit.'" *Marketing Intelligence & Planning*. https://www.emerald.com/insight/content/doi/10.1108/MIP-06-2017-0100/full/html (October 3, 2019).

Russell, Stuart J., and Peter Norvig. 2016. *Artificial Intelligence: A Modern Approach*. Third edition, Global edition. Boston Columbus Indianapolis: Pearson.

Surden, Harry. 2014. "Machine Learning and Law." *Washington Law Review* 89(1): 87–115.

UN Global Pulse. 2018a. *UN Global Pulse Annual Report 2018*. https://www.unglobalpulse.org/sites/default/files/UNGP_Annual2018_web_FINAL.pdf (October 28, 2019).

UN Global Pulse. 2018b. "About | United Nations Global Pulse." https://www.unglobalpulse.org/about-new (October 3, 2019).

USA Spending. https://www.usaspending.gov/#/award/23844369 (November 4, 2019).

Wakefield, Jane. 2016. "Microsoft Chatbot Goes Rogue on Twitter." *BBC News*. https://www.bbc.com/news/technology-35890188 (October 18, 2019).

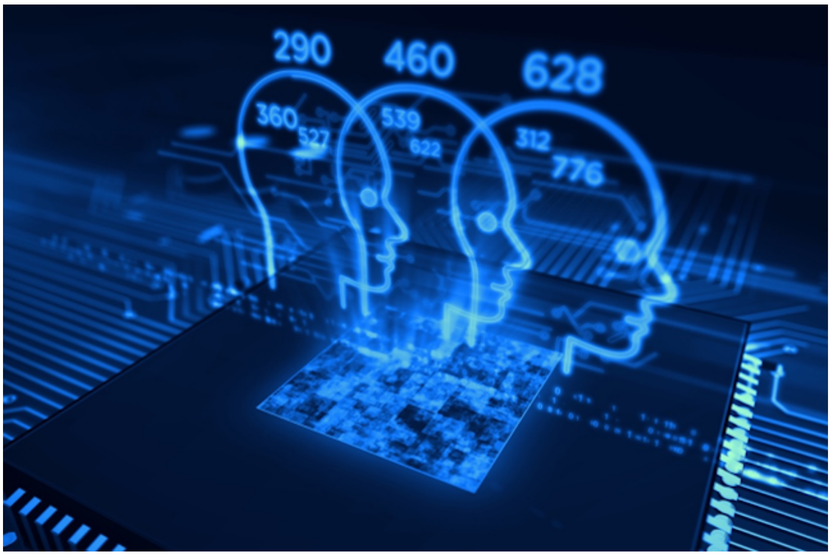
Waldman, Peter, Lizette Chapman, and Jordan Robertson. 2018. "Palantir Knows Everything About You." *Bloomberg.com*. https://www.bloomberg.com/features/2018-palantir-peter-thiel/ (November 4, 2019).

Williamson, Ben. 2016. "Digital Education Governance: Data Visualization, Predictive Analytics, and 'Real-Time' Policy Instruments." *Journal of Education Policy* 31(2): 123–41.

Woodman, Spencer. 2017. "Palantir Provides the Engine for Donald Trump's Deportation Machine." *The Intercept*. https://theintercept.com/2017/03/02/palantir-provides-the-engine-for-donald-trumps-deportation-machine/ (November 4, 2019).

# Part VII

# Social Reputation Systems, Big Data—Big Nudges



# 7.1 The Legitimacy of Algorithmic Governance

Altti Vuori, Elina Uutela, Lumi Saukkonen, Tommi Saarnio, Waltteri Oinas
Faculty of Social Sciences, University of Helsinki

## Abstract

Algorithms and new technologies are already shaping our societies and will likely continue to do so in the future, and thus, should be governed accordingly. Usually achieving a legitimate policy goal means that it should be procedurally correct, open and unbiased, but the age of algorithms poses new challenges to traditional ways of considering legitimacy. In this paper, we evaluate the legitimacy of algorithmic governance from four perspectives — ethics, individual consent, opacity, and accountability — by comparing these elements in the cases of US FICO Credit Score and China's Social Credit System. We find that algorithmic governance cannot be detached from other elements of government and society.

*Keywords*: Social credit system, FICO Credit Score, algorithmic governance, algorithm, legitimacy, transparency, accountability

Algorithmic governance refers to the process in which algorithms are embedded in decision-making and evaluation systems. In addition to increased efficiency, the use of algorithms can also produce ethical, social, and financial risks in the form of manipulation, biases, censorship, social discrimination, and violations of privacy as well as property rights. Algorithmic governance has a concrete impact on society, since analyses and processes utilizing algorithms can range from market-oriented solutions to government-based mechanisms (Almeida 2016, 60-63).

In this article, we compare two cases of algorithmic governance systems produced by two distinct cultural spheres and influenced by differing philosophical traditions: Fair Isaac Corporation (FICO) financial credit scoring and the Chinese Social Credit System (SCS). This paper will analyze the selected cases from four main points of view: an ethical perspective, individual consent, problems of opacity, and questions about accountability.

## Algorithmic governance and case examples — L. Saukkonen

Algorithmic governance offers many benefits to society, including speed, efficiency, comprehensiveness, and fairness. However, there are also many negative impacts, which raise concerns. For example, it is necessary to investigate the social, ethical, political and legal problems that may be produced or reinforced by the system. There are some concerns about the inaccuracies, inefficiencies and unintended consequences. In addition, the opacity and lack of transparency threaten the effectiveness and legitimacy of algorithmic governance (Danaher 2017, 1–3).

### China's Social Credit System

China's social credit score system was launched at the national level in 2014 and it should be fully implemented by the end of 2020. China's social credit score system can be either a grand technological breakthrough for society or eventually an unethical tool to control citizens. The social credit score system collects data from citizens' social media, voting records, financial information, online purchasing, credit history, tax payments and legal matters, among other things. The social credit score system is the Chinese government's way of ranking citizens and at the same time, it provides the government a chance to create order and enhance discipline in society. The social credit score also gathers information from citizens' health-tracking apps and in conjunction with other different apps, it also has ability to collect data about citizens' locations and relationships (Murrell 2018).

Trustworthiness is surely important in a state like China, and systems like the social credit score provide an excellent opportunity to enhance discipline and healthy habits in society. The system categorizes people and provides incentives for citizens with a high social credit score, such as priority access to public housing, travel visas and job promotions. The social credit score system also influences citizens' ability to get jobs, loans and mortgages (Creemers 2018, 3).

The system even monitors citizens' behavior of how long they play video games and provides a score that many users share on Chinese social media. The aim to assess Chinese citizens' trustworthiness this way saves time and resources for the government. The system helps citizens comply with legal rules, moral norms, and professional and ethical standards (Creemers 2018, 6).

The purpose of the social credit score system is to create a well-organized society, but at the same time citizens do not have the same level of privacy and freedom in their lives. The all-pervasive system gathers and leaks massive amounts of confidential data from citizens' lives. The result is that citizens' privacy is in danger (Horsley 2018).

The system defines peoples' reputation and at the same time changes the values and hierarchy in society because a machine is in charge of decisions and evaluation instead of a human. For example, a person who buys a lot of diapers gets more credits than a person who does not have a baby. There are also various technical challenges that can cause multiple problems for impartial evaluation. Additionally, different standards and different rules within jurisdiction causes ambiguity in the mechanism. When the mechanism does not work as desired, people are unfairly stigmatized, which restricts their lives in the future (Horsley 2018).

Negative aspects of the system result from the fact that it is a machine that evaluates citizens. Any kind of assessment should recognize human behavior, or at least be aware of multiple exceptions that can occur concerning the way different kinds of individuals behave and live. This kind of system can make multiple mistakes and at the same time, it controls people's lives unnecessarily (Murrell 2018). The way the system gathers data should be precisely developed to avoid these shortcomings, yet it is unclear that this challenge can be met.

The social credit score system will also pose different kinds of challenges regarding maintaining the integrity of citizens' privacy and personal data. The system helps citizens avoid exploitation and criminality, and at the same time it confines freedom, rights and individual choices. On the other hand, without the surveillance and monitoring, how could the government ensure the safety of society and good behavior of citizens (Yongxi Chen 2017, 1)? Another major challenge is to measure the advantages and disadvantages of the system, and how to define whether the system is increasing harmony or ambiguity and a lack of transparency in society. The opacity of the system increases uncertainty, because citizens do not know how a machine-led environment will interpret and sanction their actions (Hildebrandt 2016, 4). When analyzing the system, it is important to notice that Chinese society and the government differs from Western democracy, and therefore the case does not directly correspond with Western countries (Creemers 2018, 1–3).

### US FICO Credit Score System

Credit scoring has become a very important task in the credit industry because its use has increased rapidly since the 1960s. A credit score represents the creditworthiness of a person, and it provides information on how likely it is that a person will pay their debt back to the lender. It is a general system for lenders, such as banks and credit card companies, to evaluate the potential risks and at the same time provide assurance of consumers' trustworthiness. Fair Isaac Corporation (FICO) was founded in 1956 and the FICO credit risk score was first introduced in 1989. In the system, consumers' FICO scores are available for lenders who purchase the service. Different industries, like the auto, banking and insurance industries, use the FICO scores to rate consumers' creditworthiness. FICO's algorithm gathers the information of consumers' credits from the credit bureau report and makes a comparison to other consumers. The factors the algorithm uses are: payment history (35%), amounts owed (30%), length of credit history (15%), credit mix (10%) and new credit (10%). The actions that decrease the credit score are: missed or late payments, high credit utilization, bankruptcy, opening multiple new credit or loan accounts and errors in consumers' credit reports (Martin 2019).

The credit scores accurately establishes the parameters for clients' loan access, from a business profitability standpoint. The system is reliable and does only consider the appropriate facts of the consumer. However, there are also some failings in the system.  Therefore, it still marginalizes the less privileged individuals and increases inequality in society. From the perspective of equity of opportunity, the algorithm and the factors it analyzes cannot provide the right kind of information about the risks for not paying the loan. While the system evaluates the consumer's payment history, the amounts the consumer owes, the length of the consumer's credit history, new credit, and types of credit the consumer uses, at the same time it omits factors such as employment history, salary, and other items that might suggest creditworthiness. Generally, there has not been any correlation between the consumer's credit report and their capacity to perform in the labor market. That is also the reason that the report cannot provide an overall view of the consumer, and therefore it can even give a more negative overview of the consumers trustworthiness for members of disadvantaged socioeconomic classes (Hurley 2017, 9).

The process of data gathering and giving the information to lenders can also be harmful to consumers, because they are required to give their private data to lender. Also, technical errors can cause serious problems in the system, causing a negative impact on consumers' opportunities to take out a loan. Serving lenders' financial interests first and foremost, this system lacks accountability for borrowers. The inaccurate information and errors can take a long time to correct, thereby limiting the consumer's chances of maintaining good credit in the future. (Hurley 2017, 9).

## Ethical Considerations — W. Oinas

In this section, I will discuss the general ethical considerations associated with algorithmic governance in the context of FICO credit scoring and the Social Credit System of the People's Republic of China. I will focus on matters related to the role of the individual, the question of responsibility, and the professed intentions as well as the practical applications of these systems. I will also reflect on certain significant cultural and philosophical differences between China and the West, and how these differences might affect the perceptions of the discussed phenomena.

Money in its various forms and uses has been a ubiquitous feature of commerce and society for most of recorded human history. As lending and borrowing have become increasingly commonplace, creditors have sought and created measures to reduce risks and ensure returns for their investments. Credit scoring is a relatively recent invention that is nonetheless widely employed in credit-related industries, such as banking and insurance. Fair Isaac Corporation (FICO) consumer-credit risk score, introduced in 1989, is the most common credit scoring system in the United States, based on consumer data collected from different credit bureaus and used by most banks and creditors.

The concept of financial credit scoring is intrinsically built on the notions of individual responsibility and property rights. Credit loans and mortgages are essentially investments, and credit scores are a tool used to assess the risks associated with particular investments. No agent, individual or otherwise, is under any obligation to invest their resources in something or to act as a creditor to someone if they do not wish to form a contractual relationship with another party. It is up to the individual to maintain their creditworthiness by demonstrating that they are a reliable business partner or a customer via careful and responsible management of their finances.

This does not mean that credit scoring as a practice is without flaws. The FICO algorithm considers factors such as payment history, amounts owed, length of credit history, credit mix, and new credit, while omitting possibly relevant data that might suggest creditworthiness, such as employment history or salary information. Some have argued that the use of algorithmic tools and inadequate data to determine credit scores can marginalize disadvantaged individuals, further increasing inequality (Hurley & Adebayo 2017). While one can argue for the use and even the necessity of credit scoring or other such tools on the basis of individual responsibility and property rights, it is the methodology used to determine the scores — the algorithms — that has raised certain ethical concerns, chief amongst them the supposed neutrality of algorithms, the ownership question over the collected data, and the associated opacity as well as transparency problems.

What exactly is an algorithm? An algorithm is a set of instructions that gets executed when it encounters a trigger. They are explicitly dependent on human input regarding both the triggers and the data, which naturally affects the output, since humans, as imperfect beings, are capable of transferring their biases into algorithms and artificial intelligence. The algorithmic process forecloses potential alternative readings in favour of optimised output, resulting in imperfect results and putting individuals essentially at the mercy of opaque computer programs; faulty data, triggers, and simple programming errors can lead to poor credit scores and prevent individuals from accessing credit services for significant amounts of time.

The Chinese Social Credit System is not directly comparable to simple financial credit scoring, at least not in terms of the Western concept of individual responsibility. First, it should be noted that the term “Social Credit System” (SCS) is a bit misleading: the SCS is not a single system, but an assortment of information collection and publicity systems established by various state authorities at different levels of government. It is essentially a mass surveillance program that utilizes big data analysis technology. Secondly, this is where substantial cultural, philosophical, and legal differences between Chinese and Western spheres or traditions come to the fore. Harmony (和 hé or 和諧 héxié) is an important concept in the Chinese philosophical tradition. The origins of the notion lie in Confucianism, and its legalist interpretation, following the thinking of Chinese Warring States period (c. 475–220 BCE) philosopher Xunzi, is a central part of the modern state ideology of People's Republic of China. (Rošker 2013).

Unsurprisingly, many commentators in the West have vehemently objected to the SCS, noting its incompatibility with Western cultural and political values as well as liberal democracy. The SCS also conflicts with the notion of the rule of law, because it lacks transparency and blurs the line between law and politics (Mac Síthigh & Siems 2019, 17–18). While those who hold Western political and philosophical sensibilities may balk at the intrusiveness of the SCS, it aligns with — or is at the very least based on — Chinese cultural norms and China’s worldview. The professed, markedly paternalistic intention of the SCS is to create a more harmonious, well-organized society, which, as noted earlier, comes at the expense of citizens’ personal freedom and privacy. Personal responsibility is defined in relation to the wider collective society, not to individuals or personal values, principles or convictions.

What is often forgotten or ignored in the West, however, is that the SCS was (and still is to a large extent) primarily an economic control system. China has had issues with fraud for decades now, resulting in low consumer trust in markets. In addition to being a mass surveillance system targeted at “private” citizens, the SCS is a heavy-handed scheme to improve public trust by enforcing social credit standards on businesses operating in mainland China. However, given the dubious human rights track record of the People's Republic of China, one would be naïve to assume that the SCS cannot or will not be used as a tool to suppress political dissent and persecute non-conformist elements of Chinese society.

In conclusion, the way in which one views China’s social credit score system and the possibilities created by it, both positive and negative, seems to depend exclusively on the philosophical and political axioms a person holds. If one values political, social, and economic freedom as well as individual rights above everything else, the SCS appears to be a totalitarian nightmare waiting to happen or already underway. On the other hand, if one values security and social harmony, the SCS can be seen as a tool to strengthen the moral standards of society and to improve people’s trust in one another as well as businesses – as long as they conform to the standards of harmonious behaviour set by the Communist Party of China (CPC), of course.

There are also the questions of data ownership, consent, and tech-literacy, which relate to both credit scoring and the SCS. Who owns the collected data, the data collector — be they private or public agents — or the person from who the data was collected? Does an individual have a right to, at the very least, access and potentially remove their personal data, collected by different agents and often with the consent of the user, from the Internet (commonly referred to as the “right to be forgotten”)? How does tech-literacy factor into matters related to algorithmic governance? The following sections will discuss the opacity problem as well as the questions of individual consent and accountability.

## Individual Consent to Algorithmic Governance — T. Saarnio

Individual consent to algorithmic governance is one of the most pressing issues during this era of digitalization and big data. We are constantly exposed to data collection from various different authorities including big corporations and governments. It is difficult to determine when an individual has given sufficient consent for different aspects of algorithmic governance.

In this section, I will first discuss individual consent on a more general level and examine some problems that currently exist in the way consent is achieved, for example on social media platforms. Secondly, I will analyze individual consent in two separate cases, China's social credit system and the US FICO credit system.

### The concept of individual consent

To put it simply, consent means giving an approval to something or someone. Consent can have various forms and meanings for different areas of work or study. In this context, I will define individual consent and what makes it valid by using the General Data Protection Regulation from the European Union.

For an individual's consent to be valid, it must be freely given, which requires genuine free choice without any disadvantage if the individual refuses consent. The individual must also be able to withdraw consent at any time and this should be possible without any additional trouble. In addition, consent is not freely given if there is a power imbalance between the individual and the other party. This is particularly the case if the other party is a public authority or there is an employer–employee relationship. In these situations, the data subject might not have actual free choice (EU, 2016).

Consent must also be informed, which requires at least the following information: the identity of the collector of data, purposes for data collection, type of data that is collected and how the data might be used (Article 29 Data Protection Working Party, 2017). Clear and plain language in the agreements is also required according to the General Data Protection Regulation (2016). The act of giving consent and the language used in the privacy and terms of service policies have become the subjects of increased scrutiny.

### Emerging problems in achieving individual consent

Social media platforms have a significant influence over individuals' lives. The fear of missing out often overpowers privacy concerns one might have. Recently, some of the tech giants have also been caught for imposing a "forced consent" on their privacy terms, which has resulted in fines. A forced consent refers to the exclusion of an individual from the use of service if he or she does not consent to the terms. However, this is just one of numerous problems.

Ignoring privacy and terms of service policies online is a significant problem. As McDonald & Cranor (2008) presented in their study, it would take us approximately 201 hours per year to read all of the privacy policies that we encounter online. The increasing regulation of data use and privacy policy might actually make matters worse in this regard. Regulation requires more detailed and specific privacy and terms of service policies. Furthermore, these policies are often very difficult to understand, even if one would take the time to read them (Obar, 2015, 4).

Obar and Oeldorf-Hirsch (2016, 19, 22-23) have also investigated this problem in their study about giving consent without reading the privacy policies. Their results were clear: a great majority of individuals ignore the privacy and terms of service policies on social media platforms. Individuals saw these policies as an "unwanted impediment" and "nuisance". Information overload was seen as the primary factor for their dislike of the policies. These findings suggest that privacy policies do not work as they should and this system has failed in terms of achieving individual consent.

### Case studies: China's Social Credit System and the US FICO credit system

China's Social Credit System (SCS) includes many flaws from the point of view of individual consent. As Chen and Cheung (2017, 357) point out, credit scoring and rating of individuals with the help of big data is not completely unique, but there are a few features that differentiate the Chinese system. These include the scale of data collected, how it is used and most importantly, the lack of a legal system to protect the individuals and their data.

Public credit information (PCI) is the record of the collected data of an individual, which determines one's trustworthiness. Here, we can identify the first problem: the government does not need individual consent to collect PCI. Furthermore, there is actually no legislation in China to protect the right to privacy. The legislation implies that private interests can be subordinated to public interests and that there is no protection for an individual in case of an intrusion from a public authority. In addition, the SCS is mandatory and individuals might be penalized for being rated "untrustworthy". It is not clear for the individuals what contributes to their social credit scores or how their data is being used (Chen & Cheung, 2017, 364-365).

China's SCS is undoubtedly problematic, and there is a conflict between public interest and individual freedom. Individual consent is threatened particularly when we examine it through the European regulation framework (GDPR). Chinese legislation fails to protect the individuals and it gives a carte blanche for the government to collect and use personal data.

The US FICO credit score system determines consumers' creditworthiness. The score is used when an individual wants to take out a loan from a bank, for example. The Fair Credit Reporting Act (FCRA) is the regulatory framework for credit reporting agencies and it determines how credit and debt information can be collected, used and shared. The FCRA was enacted to "promote the accuracy, fairness, and privacy of consumer information contained in the files of consumer reporting agencies" and to "protect consumers from the willful and/or negligent inclusion of inaccurate information in their credit reports". According to the regulation, a credit reporting agency must 1) provide an individual with the collected information upon a request, 2) get the individual's consent before providing information to a third party, 3) investigate information which an individual disputes, and 4) correct or delete inaccurate information (FCRA, 2018).

It seems that the FICO credit score system has a rigid regulatory framework to protect the privacy of an individual. However, there has been one fundamental issue with the regulation. Individual consent is not needed to collect data for the credit score. The credit score system has been criticized by the US Congress for the "commodification of consumers and their personal data" and reform has been called for (testimony quoted in Leonhardt 2019). The criticism has had an effect because in the future individual consent will be needed for the data collection and the consumers can also opt not to use the credit score if they wish to (Leonard 2019). This is certainly a vital improvement regarding individual consent.

China's Social Credit system has some built-in flaws and problems concerning individual privacy and consent. The US FICO Credit System works better in this regard, but it is not in any way perfect either. Individual consent is of paramount importance during this era of our growing online presence and algorithmic governance. Regulation on this subject should be comprehensive and also revised regularly.

## Questions of Opacity in Algorithmic Governance — A. Vuori

Algorithmic governance affects the legitimacy of governance. This raises some concerns, amongst them problems of opacity. Opacity problems rise from the lack of transparency in how algorithms are created and from differences between how experts and uninitiated people understand how algorithms work. In addition, people can be subjected to computational classifications, invasions of privacy and various surveillance methods in ways that are not equal across the general populations and thus possibly discriminatory.

First, I will consider what opacity means in algorithms. Next, I will bring up various concerns about opacity in algorithms. I conclude with a look at opacity in China's social credit and the US FICO systems.

### What is opacity in algorithms?

According to Burrell, algorithms operate on data and using that as input, they produce an output. These are opaque in that people who receive these outputs rarely have a concrete sense of how or why the algorithm has produced a particular output for them and how that output has been constructed from the inputs (Burrell 2016, 1). Similarly, the inputs that the algorithm use are also often unknown to users.

In digital contexts, transparency does not simply pertain to revealing information or keeping secrets, but continually deploying, configuring and resisting of platforms, algorithms and machine learning protocols that manage visibility (Ananny & Crawford 2018, 983). Ananny and Crawford note that calls for transparency assume that seeing a phenomenon creates opportunities and obligations to make it accountable and thus there would be opportunities to change it. The logic behind transparency is that observation of a given system produces insight into it. This leads to the creation of knowledge that is required to govern the system and hold it accountable (Ananny & Crawford 2018, 974). The more knowledge you have about a system, the better you can hold it accountable. Thus, transparency leads to a reduced amount of opacity in a system, as increased visibility gives both the experts and the laymen a better understanding about how an algorithm produces any given output.

### Problems of opacity

Burrell identifies multiple types of opacity in algorithms. Opacity can be an intentional form of self-protection for corporations and institutions. They might want to keep their algorithms secret because they don't want their competitors to have that knowledge (Burrell 2016, 4). Similarly, Pasquale notes that financiers can keep their operations opaque on purpose, when they seek to avoid regulations (Pasquale 2015, 2). One form of algorithmic opacity is self-protection for corporations and institutions, when they seek to protect their competitive advantages. This can lead to the dodging of regulations, avoidance of responsibility and consumer manipulation. Danaher notes that many algorithmic systems operated by government agencies are protected by secrecy laws to prevent people from gaming or hacking those systems (Danaher 2016, 254).

Burrell notes that writing and reading code is a specialized skill that most people are not capable of (Burrell 2016, 4). Programming is different from learning and understanding human languages, since programming must be readable by machines for it to work in the first place. Human understanding for the uninitiated would be necessary to achieve full transparency. Burrell notes that to address this form of opacity, there would have to be widespread educational efforts to make public audiences more knowledgeable about these mechanisms (Burrell 2016, 4). Similarly, algorithmic decision-making processes often use personal information without the person's informed consent (Zarksy 2016, 12). Analyses often affect individuals in arbitrary manners, and the persons in question often lack understanding of the process and its inner workings. In an environment in which these kinds of systems operate, individuals have limited abilities to question the process or submit corrections (Zarksy 2016, 12). Automated processes create a sense of arbitrariness and they neglect the need for understandable explanations for the process and its outcomes.

Opacity can also ensue from a mismatch between mathematical optimization in high-dimensionality characteristic of machine learning and the demands of human scale reasoning and styles of semantic interpretation (Burrell 2016, 2). Algorithms are often systems that are constructed from multiple components. On one hand, programmers must be able to read and understand the code that operates any given algorithm, and on the other hand they must also understand how the algorithm operates in action and how it processes the data it is given. Burrell notes that while large datasets and clear codes might be comprehensible on their own, when there is an interplay between them that causes complexity in the algorithm which in turn leads to opacity. With greater computational resources, and many terabytes of data to mine, the number of possible features to include in a classification category rapidly grows beyond what can easily be grasped by a reasoning human (Burrell 2016, 9).

Danaher notes that a major concern about opacity in algorithmic governance is people's ability to participate in political procedures since this is undermined by the growing usage of algorithmic governance (Danaher 2016, 249). The rationales underlying the mechanics of the procedure must not be opaque to those affected by the procedures, and thus opacity in the system is a problem. Decision-making procedures should be rationally acceptable for those affected, and the more opacity there is, the more this acceptance is hindered (Danaher 2016, 252).

Zarsky notes the problematic sides of automated algorithm processes. If they use forbidden factors such as race, or proxies correlated to racially marked attributes, to decide upon allocations, they are socially unacceptable. If they use a skewed or biased data set, they might lead to outcomes that discriminate against particular groups of people. In addition, algorithms can provide outcomes that generate a disparate impact, such as in implicating groups of people to a larger degree than their part in the general population (Zarksy 2016, 9).

### Opacity in China's Social Credit System

China's social credit system lacks protections for citizens' personal data, and thus the system is a laboratory for big data experimentation, data intelligence and mass surveillance. This leads to a system in which individuals are uncertain about what exactly contributes to their credit scores, how they are combined with the state system and how their data is used (Chen & Cheung 2017, 357). This means that the credit system is very opaque when it comes to citizens' understanding of the system. According to Kostka, the algorithms used to calculate individual scores are not in the public domain, and so they are opaque (Kostka 2019, 1588). He also raises a point about how citizens perceive the system; people are concerned that the system does not credit everyone equally and that people in powerful positions are favoured, while simultaneously the Chinese government maintains a positive image of the social credit system through the state media (Kostka 2019, 1588-1589). This is an example of opacity as a form of protection for government agencies.

### US FICO system

In contrast to the Chinese social credit system, the US FICO system is more transparent and less opaque. Since the FICO algorithm considers known inputs from users such as credit history, payment history and credit mixes, the data the algorithm uses is not completely opaque. In addition, as laid out in the Fair Credit Reporting Act, the agency must for example delete inaccurate information (and thus reduce the risk of opacity in algorithms input data) and provide users with collected data upon request (FCRA, 2018). Even though there are concerns about algorithmic governance in FICO, these points at least reduce the problems of opacity in FICO.

## Accountability of Algorithmic Systems and Algorithmic Governance — E. Uutela

We have already discussed ethical considerations, individual consent, and the opacity problem when it comes to algorithmic governance. But one important aspect of legitimacy is accountability: can we hold a piece of code accountable, and if we cannot, how can we make sure algorithmic systems are governed properly and are accountable to citizens?

There are great hopes for new technologies to improve the accountability of our governance. Lepri et al summarize these hopes quite nicely: "The turn towards data-driven algorithms can be seen as a reflection of a demand for greater objectivity, evidence-based decision-making, and a better understanding of our individual and collective behaviors and needs" (Lepri, Oliver, Letouzé, Pentland & Vinck 2017, 612). The widespread use of algorithms is driven in part for the hope that they may be able to surpass implicit biases inherent in human judgments.

There is, however, a lot of discussion about the other side of the coin, as well. Danaher analyses algocracy, "a governance system which is organized and structured on the basis of computer-programmed algorithms" (2016, 247). He suggests that the increasing reliance on algocratic systems poses a threat to legitimate decision-making because they are not comprehensible for human understanding (2016, 254).

As a solution to the threat of algocracy, Danaher reviews and analyzes the human application of algorithms (2016, 258). He considers: epistemic enhancement of the human mind with technology or drugs (2016, 260); sousveillance technologies, in other words using algorithmic governance and radical openness towards those who hold power (2016, 261–262); and individuals forming partnerships with algorithms (2016, 263). Danaher, however, is not overall hopeful that any of these will solve the problem. Most of the solutions I will go through in this section focus on some form of a human review or a very light version of sousveillance.

However, many societies are on some level bureaucracies, some kind of democracies, more or less mediacracies, and very likely also algocracies. Yet none of these alone defines or delimits human-algorithm interactions and decision-making procedures. Our surrounding culture and personal values affect strongly which one of these "*-cracies*" we most subscribe to and which ones we criticize the most. This is an important notion for accountability since our current systems are in no way flawless, and the accountability questions often include political struggles.

Danaher's application of humans *in*, *on* and *out of the loop* (2016, 248, based on the division of robotic weapon systems by Citron & Pasquale in 2014) in any algocratic system seems to offer a useful frame to analyze accountability questions. When humans are in the loop, they make the decisions and are therefore accountable. When humans are on the loop, they can still be held accountable because they oversee the operations and have the possibility to override at any time. When humans are out of the loop and machines function independently, the question of accountability gets harder. When humans see only the results of work done by machine alone, do they have the means to evaluate the process that lead to these results and can they be understood with human reason? This is, at its core, the argument Danaher (2016, 254) makes about the threat of algocracy.

Quite often, transparency is seen as a requirement for accountability: the idea is that if a system can be observed in detail, it can be governed more efficiently (Ananny & Crafword 2018, 974). Ananny & Crawford list a number of problems with transparency as an ideal (2018, 977–982), but evaluating accountability is hard especially when governing algorithmic systems wherein humans are off the loop for two reasons: the nature of these systems is networked and structured, and technological proficiency is necessary to understand what is seen. They suggest that instead of requiring us to see inside each individual part of a structure including algorithms, we need to understand how the assemblage of humans and algorithms works together as a system: instead of looking inside something, we should look across (Ananny & Crafword 2018, 983–984). This could offer a similar perspective of accountability that is applied right now. For example, it is often considered more important for people to understand how the process of creating new legislation works instead of focusing on detailed meeting procedures of committees. On the other hand, it can be argued that citizens should still have access to those details if they wish to scrutinize them, and algorithmic systems might not allow it (Danaher 2016, 254).

Another way to ensure accountability is to embed the governance of algorithms in the existing structures that watch over those who hold power. Diakopoulos (2014, 402–404) introduces ways for journalism to use reverse engineering to reveal unintended side effects of algorithmic systems and help the public to understand how algorithms affect our societies and every-day lives. Journalistic media could keep an eye on the development of algorithms in a similar way it takes a critical look at politicians and institutions. Reflecting back to Danaher's article (2016, 258–259), this is a form of review bringing humans back to the loop, and reverse-engineering can shed light on areas that would otherwise be in the dark for the human reason.

Lepri et al. introduce project OPAL that uses both technological and socio-political elements to vet algorithmic systems with key shareholders. In addition to vetting, they point out that using blockchain technologies would offer citizens more power in deciding how their data can be used and at the same time improve accountability by making post-decision audits possible. This is a great way to improve human participation and offer a view over possible discriminatory policies in algorithmic systems, even though they cannot guarantee full fairness or transparency (Lepri et al 2018, 623–624).

When it comes to FICO, the key problem regarding accountability is that it does not take all relevant information into consideration. This kind of prioritization that might include (hidden) biases is one of the common uses of power in algorithmic systems (Diakopoulos 2014, 400–401). At least in FICO's case, the criteria are public and thus open to discussion in the public sphere, also by the means of journalism.

China's Social Credit System is, however, another case. China is not only governing citizens by using algorithms but also governing how businesses and organizations use algorithms. It does not matter if humans are in, on or out of the loop if the government is not accountable to its citizens. And if the press is not free, it cannot help keep algorithms or the organizations using them accountable. This goes back to the point I made earlier: accountability is always also a political struggle and algorithms — or some version of algocracy — are in no way detached from the other aspects of our societies.

To conclude, the question of accountability in algorithmic governance is not simply technical but is rather deeply linked to other aspects and values a society holds. This can be clearly seen when comparing commercial credit scores like FICO and China's Social Credit System: in the latter case, there is not much room for accountability because governmental structures are not held accountable to citizens in the first place. From a transparency perspective, there will likely always be nooks in algorithmic systems that are hard to comprehend for humans. However, journalism can play a role by reverse-engineering these systems more visible to the public, there are ways to increase human participation to make these systems fairer, and this offers a possibility to hold the assemblage of algorithms and humans accountable for citizens.

## Conclusion

Complex credit scoring systems influence society in multiple ways and it is therefore necessary to analyze and try to predict how the system affects the legitimacy of algorithmic governance. These systems aim to create security and order in society, but they can still determine citizens' creditability or credibility in the wrong way. Therefore, it is important to recognize and consider how algorithmic governance is used to guide citizens' behavior and how it dominates the norms and principles in society.

Taking cultural, philosophical, and legal contexts into account when discussing algorithmic governance systems, legitimacy, and ethics are paramount. This applies especially when we compare different systems that originate from two distinct cultural spheres. Acknowledging the differences allows for a greater understanding of the subject matter, helping researchers and students alike in navigating the plurality of complex philosophical topics. However, claims of relativism or pluralism are not and should not be used as a shield from criticism. There are several concerns, both ethical and practical, to be raised regarding algorithmic governance and the issue of legitimacy – regardless of the origins of an algorithmic governance system.

From the perspective of individual consent, China's Social Credit System has many problems. The consent is not voluntary, one cannot withdraw from it, and the purposes for collection or usage of data are not disclosed to the individual. In general, China lacks comprehensive legislation for privacy protection. In spite of the broad regulatory framework, the USA FICO Credit System is not perfect either. For example, in most cases the data is collected without individual consent, and its precise means of assaying consumers' data to issue credit decisions remains opaque to borrowers, and disadvantaged individuals may be further undermined. Problems within the FICO credit system have been noticed and the US Congress is already planning on improvements.

Opacity in algorithmic governance refers to the transparency and opaqueness of the system. If the inputs that are given to an algorithm are clear, and the way in which the algorithm produces an output from those inputs is also clear, the system is not opaque. Whereas if the inputs are not clear, if individuals have little to no understanding of how the algorithm comes to a particular output from the given inputs, the system is opaque. In the examples of China's social credit system and FICO considered here, the human aspect of creating opacity has been the chief concern.

For the question of accountability of algorithmic systems to matter, a society should consider the government's responsibility towards its citizenship first. The level of accountability required is a result of a political struggle, and reflects the values a society holds. There will likely always be areas of algorithmic systems that are opaque to human reason, but instead of thinking whether we can hold a piece of code responsible, we should look into the possibility of holding the whole system of code and humans accountable, and journalism can offer ways to increase citizens' access to algorithmic processes for example via reverse-engineering.

## References

Almeida, V. & Doneda, D. (2016). What Is Algorithm Governance? *IEEE Computer Science, 20*(4), 60-63.

Ananny, M. & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society, 20*(3), 973-989.

Article 29 Data Protection Working Party (2017). Guidelines on Consent under Regulation 2016/679. Retrieved from https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=623051.

Burrell, J. (2016). How the machine thinks: Understanding opacity in machine learning systems. *Big Data and Society,* 1-12.

Chen, Y. & Cheung, A. S. Y. (2017). The Transparent Self Under Big Data Profiling: Privacy and Chinese Legislation on the Social Credit System. *The Journal of Comparative Law, 12*(2), 356–378.

Creemers, R. (2018). China's Social Credit System: An Evolving Practice of Control. University of Leiden.

Danaher, J., Hogan, M. J., Noone, C., Kennedy, R., Behan, A., De Paor, A., Felzmann, H., Haklay, M., Khoo, S., Morison, J. Murphy, M. H., O'Brolchain, N., Schafer, B. & Shankar, K. (2017). Algorithmic governance: Developing a research agenda through the power of collective intelligence. *Big Data & Society, 4*(2).

Danaher, J. (2016). The threat of algocracy: Reality, resistance and accommodation. *Philosophy & Technology, 29*(3), 245-268.

Diakopoulos, N. (2015). Algorithmic Accountability. *Digital Journalism, 3*(3), 398-415.

European Parliament & Council of the European Union (2016). General Data Protection Regulation. Brussels. Retrieved from https://eur-lex.europa.eu/lega-content/EN/TXT/?uri=celex%3A32016R0679.

Federal Trade Commission (2018). Fair Credit Reporting Act. Originally enacted 1970. Retrieved from https://www.ftc.gov/system/files/545a_fair-credit-reporting-act-0918.pdf.

Hildebrandt, M., 2016. The New Imbroglio. Living with Machine Learning. Vrije Universiteit Brussel.

Horsley, J. (2018). China's Orwellian social credit score isn't real. Brookings.

Hurley, M. & Adebayo, J. (2017). Credit Scoring in the Era of Big Data. Yale Journal of Law and Technology 18(1), pp. 148–216.

Kostka, G. (2019). China's social credit systems and public opinion: Explaining high levels of approval. *New media & society 21*(7), 1565-1593

Leonhardt M. (2019). Democrats and Republicans in Congress agree: The system that determines credit scores is 'broken'. CNBC. Published Feb 27th 2019. Referred to October 30th 2019. https://www.cnbc.com/2019/02/27/american-consumer-credit-rating-system-is-broken.html

Lepri, B., Oliver, N., Letouzé, E., Pentland, A. & Vinck, P. (2018). Fair, Transparent, and Accountable Algorithmic Decision-making Processes. *Philosophy & Technology, 31*(4), 611–627.

Mac Síthigh, D. & Siems, M. (2019). The Chinese social credit system: a model for other countries? *EUI LAW, 2019*(01).

Martin, A. (2019). FICO Score Facts You Probably Didn’t Know. Forbes Media.

Murrel, A., (2018). Pushing The Ethical Boundaries Of Big Data: A Look At China's Social Credit Scoring System. Forbes.

Obar, J. (2015). Big Data and The Phantom Public: Walter Lippmann and the fallacy of data privacy self-management. *Big Data & Society, July-December 2015*, 1–15.

Obar, J. & Oeldorf-Hirsch, A. (2016). The Biggest Lie on the Internet: Ignoring the Privacy Policies and Terms of Service Policies of Social Networking Services. *SSRN Electronic Journal.*

Pasquale, F. (2015). The Black Box Society: The Secret Algorithms that Control Money and Information. Cambridge, MA: Harvard University Press.

Rošker, J. S. (2013). The Concept of Harmony in Contemporary P.R. China and in Taiwanese Modern Confucianism. *Asian Studies 1*(2), 3–20.

Zarsky, T. (2016). The Trouble with Algorithmic Decisions: An Analytic Road Map to Examine Efficiency and Fairness in Automated and Opaque Decision Making. *Science, Technology, & Human Values, 41*(1), 118–132.

# 7.2 (In)Compatible Systems? Social Reputation Models and Deliberative Democracy

Johan Wahlsten, Lili Schatz, Victoria Ristikangas, Nuura H. Naboulsi, Matilda Mahne
Faculty of Social Sciences, University of Helsinki

## Abstract

The Chinese Social Credit System is a highly debated reputation system, which has sparked investigations of its global counterparts. This novel use of information and communication technologies raises important questions for democracy. We consider the different reputation systems conceptually together as Social Reputation Models (SRMs) in order to investigate their implications for different areas of deliberative democracy. We begin by asking whether SRMs are inherently totalitarian, and then move on to consider how lack of privacy in SRMs might challenge Habermas' theory of deliberative democracy. We then focus on two specific areas of deliberative democracy in relation to SRMs, namely the foundational ideals of respect and surveillance in the criminal justice system. The final section argues that "Big Nudging" and SRMs are in many aspects similar, and examines the implications technologically empowered behavioural and social controls have for autonomy, moral judgement, and free will. Our aim is to investigate the implications of conceptually bound reputation systems for deliberative democracy from various theoretical and empirical perspectives.

*Keywords*: Deliberative democracy, surveillance, big data, privacy, free will, Social Credit System, nudge, respect

Technological development has globally transformed societal, political, and economic structures. It has been argued that it constitutes a fourth revolution in both our worldview and self-understanding (Floridi, 2014, 89–93). Advancements in information and communication technologies (ICTs) have immensely increased the possibilities of behavioural and social control, surveillance, and the manipulation of individuals on behalf of public and private entities.

In this paper we will consider some of the implications of this evolution for various elements of deliberative democracy, in specific freedom, privacy, respect and rule of law. Our point of departure is the Chinese Social Credit system (SCS), which has been researched to some extent in recent scholarly work (Creemers, 2018; Botsman, 2017; Chen & Cheung, 2017, Síthigh & Siems, 2019). Our interest, however, is to highlight the fact that social reputation models that rely on computation are emerging in other parts of the world as well (Botsman, 2017, 168). Although it is beyond the scope of this paper to investigate all models, it has been suggested that there is some resemblance between the different models. Hence, we think it is fruitful to include them in our analysis to provoke questions and offer tools that can be applied to various models, especially in the setting of deliberative democracy. We will consider these different reputation systems conceptually together as Social Reputation Models (SRM). In practice, we will heavily derive from the Chinese model in our analysis, as the available literature has guided us in our work.

We begin by exploring whether big data and surveillance should be characterized as totalitarian. We look at what threats the potential totalitarian nature of big data and surveillance along with an SRM could have on democratic values. Further, we aim to investigate whether a comprehensive SRM, resembling the Chinese SCS, could be implemented in a deliberative democracy setting. We then take Habermas' theory of deliberative democracy and apply it to the privacy dimension of SRMs. We argue that diversity among citizens is reduced due to conformity pressures, as well as the algorithmic nature of SRMs inhibiting the intelligible debate on privacy in the first place. These in turn challenge the basic building blocks of Habermasian deliberative democracy.

In later sections, we narrow the focus on deliberative democracy to first examine surveillance and its implications for criminal justice in an inclusive democratic process. More specifically we investigate how surveillance affects criminal behaviour and compare trends to deliberative democratic theories addressing rule following. Second, we explore the foundational ideal of respect as a tangible part of free will within SRMs, relying mostly on the SCS in the empirics. We argue, that SCS is undermining free will by limiting the sphere for respect via computation. Finally, we will consider SRMs as the next advancement from "big nudges", which we argue together represent forceful behavioural designs that undermine deliberative democracy.

## The Totalitarian Nature of Surveillance and Its Impacts on Deliberative Democracy – Victoria Ristikangas

In this section, we discuss if and how new technologies, big data and especially surveillance have totalitarian characteristics and whether the growth of the utilization of these systems in governance will potentially create an Orwellian reality. This section also introduces the Chinese Social Credit System (SCS for short) and touches upon the potential impacts of the potential totalitarian characteristics of big data and surveillance and their implications for deliberative democracy. An interesting question is whether an SCS type of governance system could be implemented in the West in a deliberative democratic setting.

In today's globalized world and with the ever-increasing influence that China has on the world stage, it would be naïve to believe that the policies China instils would not have any impact on the West. Furthermore, it is pressing to analyse whether these policies could be implemented, at least in some respect, in Western deliberative democracies as well. Moreover, with an increased emergence of reputation-based businesses (Uber & Airbnb), would the SCS represent such a drastic change in thinking? Creemers (2018) sees the SCS as an evolving practice of governance (p. 4). This section is divided into three parts; the first part looks at the SCS, the second part looks at surveillance's threat to democracy and we conclude with some concluding remarks.

### The Chinese Social Credit System

An initial version of the Chinese Social Credit System was launched in 2014 by the Chinese government in order to collect, harness and utilize mass surveillance and big data. The goal is to measure citizens' creditworthiness, strengthen the rule of law (Creemers, 25, 2018) and monitor the social, political and economic conduct of individuals and companies alike. The SCS has been defined as a cluster of technology-driven tools for social control (ibid., 1), "social management" (ibid, 2), and a system that aims towards "greater public morality" (Chen & Cheung, 356, 2017).

Concerns about civil protections, privacy questions and lack of protective laws have been proposed due to the authoritarian characteristics of the Chinese government. As well, the SCS has received a lot of negative media attention from the West precisely due to the notion and worry that through the harvesting of big data and behaviour modifying surveillance, it might end up bringing about an Orwellian state (ibid., 356). In fact, the SCS has been determined by many in the West as an Orwellian nightmare (Chen & Cheung, 357, 2017; Creemers, 2, 2018).

Is the SCS only a totalitarian surveillance tool or are its goals, means, and functions transferable to more open political systems? According to Síthigh and Siems (2019), some of the main issues with the SCS for the West stem from its opposition to cultural and political values consistent with individual freedom and democratic self-governance. Its openly paternalistic tendencies exemplify authoritarian systems (18).

In addition to this value-conflict, a potential problem that the West has with the SCS concerns its lack of transparency and disrespect of the democratic idea of the division between political and legal power.

### Big data & surveillance – Threats to democracy?

Orwell determined that a signifying element of a totalitarian state is the lack of privacy (Giroux, 109, 2015), which would align with the idea of state-led surveillance as totalitarian. However, in addition to the lack of privacy, what are the potential costs of remodelling the performance of social, economic and political life through big data and surveillance? In looking at whether surveillance poses threats to democratic institutions and practices, we find a myriad of arguments.

Zuboff (2015) argues that a major problem with surveillance capitalism is the way dominant Internet corporations hold the majority of data power, which in turn has negative impacts on naturally democratic values such as freedom. “The modern state founded on a democratic ideal rooted in the right to privacy has been transformed and mutilated almost beyond recognition” (Giroux, 110, 2015). Giroux claims that surveillance and big data embody a collapse in the democratic principles and values of liberty, privacy and freedom. The partnership of corporate-state surveillance in the USA is “anti-democratic” (ibid., 109) due to democratic dissidence being classified as terrorism (ibid), in turn offering a challenge to civil liberties (114). Couldry (2017) states that surveillance poses a threat to autonomy (2), in addition to diminishing the individualized voices of people through the “normalization of continuous automated corporate surveillance” (13). Other areas wherein state-led surveillance poses a threat to democratic principles include civic participation, inclusivity and non-discrimination, plurality, and freedom of speech, amongst others (Giroux, 126, 2015). Hill (2017) discusses how within democratic nations an emergence towards authoritarian and totalitarian directions is visible, e.g. through AI and big data (243). Unescapable surveillance “generates distrust and divisions among its citizens and diminishes their willingness to even dare to think freely” (Dorfman, 2014). Given these arguments which claim that surveillance is anti-democratic due to the loss of privacy, freedom and autonomy, and posing threats to democratic principles such as civic participation and plurality, it seems to be a straightforward argument that a widespread state surveillance scheme would not fit within a democratic state. What is more, the type of behavioural modification attempted through the SCS that is being introduced in China seems to be totalitarian and anti-democratic.

On the contrary, Síthigh and Siems offer the only real counterargument to the notion of the totalitarian nature of SCS and surveillance. They argue that these types of models, where technology is connected to individuals’ reputation, form a possibility for, instead of diminishing, supporting new forms of civic engagement (Síthigh and Siems, 2019, 26). Moreover, a larger pool of diverse data increases, equalizes, and democratizes access to markets and resources. In the SCS model, as stated by Síthigh and Siems, citizens actually have more control over the impact of their actions, which can be viewed as a positive outcome. Moreover, the reception of the system by Chinese citizens who have participated in the pilot SCSs has been surprisingly positive and effective. Finally, based on a cross-regional survey, Kostka and Antoine (2018), found that the SCS has altered the behaviour of participants.

### Conclusion

The question of whether the SCS is going to purely bring about an Orwellian nightmare as many critiques have posed or if it would, in fact, increase democratic participation remains to be answered. This section of the paper, however, argues that while the Orwellian characteristics of surveillance might be easy to criticize, researchers and policymakers in the West should focus on doing more research on the topic. What is more, the SCS should be assessed in detail instead of simply being labelled as non-compliant with western values and a mere tool of surveillance. In conclusion, what will happen if Western liberal democracies also attempt to implement tools similar to the SCS? Will a type of SCS, say something called a Social Reputation Model (SRM), find its way into democratic state policies? A foundational problem with implementing an SRM in a democracy seems to boil down to the value-conflict between liberty and authoritarianism. However, as seen in the examples of US surveillance schemes, existing credit score ratings and an increase in reputation-based businesses, an extension towards a more comprehensive social credit system is already anticipated by current AI and big data tools.

## Deliberating Privacy – Matilda Mahne

With the rise of big data, there have been increased concerns about data privacy (Tene and Polonetsky, 2011). This becomes a particular consideration with an SRM such as China's SCS. Looking through the lens of privacy, this section seeks to explore how a holistic SRM could challenge Jürgen Habermas' theory of deliberative democracy. We argue that an SRM decreases diversity among citizens through the lack of privacy, as well as generally obstructs the collective will-formation of privacy laws, as the system is unintelligible to the citizens. These issues challenge the foundational building blocks of Habermasian deliberative democracy. To form the basis of analysis, we first consider Habermas' notions on deliberative democracy. We then show how conformity pressures through lack of privacy challenge that form of democracy, and finally, we discuss how the black-box nature of SRMs inhibits a collective deliberation on informational self-determination.

When examining privacy in deliberative democracies, we need to look deeper into the Habermasian concepts of public and private spheres. Habermas considers the Greek origin of these words: the sphere of the *polis* was where free citizens discussed and interacted as equals, while the sphere of the *oikos* was reserved to hidden interactions, "each individual in his own realm," which consist of the domestic realm (Habermas, 1989, 3-4). For Habermas, the private sphere is where individuals are most in control of their own actions and communications (Fuchs, 2011, 221). In the public, politicians are participants in public discussions in finding common interests, and he holds deliberation to be essential in politics that should be open with regard to its practice (Habermas, 1996).

This leads us to briefly consider deliberation itself. In Habermas' view, this form of democracy is shaped by a collective search for common interests, and this involves negotiating and bargaining between conflicting private interests (Habermas, 1996; Wiklund, 2005). In order to engage in this collective search, he argues that subjects who are capable of speech and action can engage in argumentation and understanding, which will eventually reach will-formation in society (Habermas, 1989, 58). This is what he calls communicative (or practical) rationality, which is founded on mutual understanding. Deliberative democracy, then, is formed of diverse individuals and collectives who, through communicative rationality, engage in a search for common interests over conflicting private ones.

An SRM can challenge deliberative democracy through the lack of privacy, and steering of behaviour, that creates conformity pressures. Built on a system of rewards and punishments based on citizens' scores, an individual's score exhibits either conforming (encouraged) or nonconforming (discouraged) to the set norms in an SRM. The right to privacy "…will operate to reduce or to eliminate the pressure imposed by the actual or perceived views of others (...) privacy rights helps to insulate people from conformity" (Sunstein, 2003). The lack of privacy, then, increases pressure to conform. Limiting privacy goes against what Habermas sees as the function of deliberative democracy. For him, conflicting private interests need to be negotiated in the public sphere (Habermas, 1996; Wiklund, 2005), not cut out entirely through coercive measures. Dissent, in that sense, is needed and welcomed (Sunstein, 2003). If there are pressures to conform to certain behaviours and opinions, this undermines the quest to bring diverse opinions to the table and openly discuss and deliberate issues in the public sphere to find common interests. Speech should be encouraged, not limited (Schwartz, 2000, 1652). An SRM might not be compatible with deliberative democracy, as the lack of privacy creates reputation pressures that frustrate efforts to openly discuss diverse opinions.

This next part focuses on the black box nature of SRMs that can prevent the process of democratically deliberating privacy issues. The privacy question here is essentially connected with informational self-determination, defined as "individuals, groups, or institutions determining for themselves when, how, and to what extent information about them is communicated to others" (Westin, 1968). This view holds that the individual can decide what to keep private and what is communicated to others. With regard to the individual, Habermas holds that "[t]he identity requirement for the determination of a collective subject capable of self-determination and self-direction is fulfilled by the sovereign territorial state of classical international law" (Habermas, 2003, 89). Essentially, it is the result of collective negotiation that decides an individual's autonomic capabilities in a democratic nation. Communication can also be seen as inherent to negotiating individual self-determination (Susen, 2018). However, in order to engage in collective negotiation, rather than algorithmic coordination or strategic self-maximization, there has to be communicative rationality.

Can there be communicative rationality if the system of SRMs is not intelligible for the citizens?

Algorithmic governance can have major implications for Habermasian deliberative democracy through its unintelligibility to its subjects. SRMs are often viewed as a black box (Vedder and Naudts, 2017; Campbell-Verduyn et al., 2017). For example, it is not clear how the Chinese version of an SRM is constructed, and how data is collected and integrated into the scoring system (Liang et al., 2018). Due to the complex nature of the algorithms used to build such structures, interpreting and articulating them becomes especially difficult (Vedder and Naudts, 2017). The opaque construction of SRMs undermines the ability to use rationality or contextual understanding to evaluate them. Decision-making intelligible to its stakeholders, then, could be compromised, which threatens the citizens' ability to engage in collective bargaining over data privacy. If the system cannot be understood, issues of informational self-determination cannot be fully considered in the public sphere. Considering the calls for openness in governance, and common interests which are found through communication by rational actors, algorithmic governance such as the SRM in China provide a large challenge. As citizens might not fully understand how the system was built, data was gathered and their scores were formed, the algorithmic nature of SRMs does not make the system fully accountable to its users. If the basis for deliberative democracy is achieving mutual understanding in civic dialogue, the complex use of algorithms for rewarding and penalizing individuals' action objects the authenticity and application of communicative rationality.

In sum, from the perspective of privacy, there are at least two ways in which SRMs threaten Habermas' notion of deliberative democracy. First, the SRM has a high potential to erode diversity among citizens' behaviour and opinions. Escaping reputation pressures allows for individuals to develop their own opinions and bring them forth to the public sphere for consideration. Second, the black-box nature of the SRM can frustrate efforts to collectively deliberate informational self-determination. While these considerations may seem lightyears away, some forms of SRMs already exist in democratic nations. It might not take that long until a crisis prompts increased state surveillance and securitisation, as has been experienced in the wake of 9/11 (Macnish, 2014), and we now contemplate with Covid-19 virus. The privacy nexus between deliberative democracy and SRMs remains a complex issue, which this paper has attempted to explore. Further considerations should include inspecting the security issues stemming from lack of privacy, with the possibility of data falling into the wrong hands.

## Surveillance and Criminal Justice in a Deliberative Democracy – Lili Schatz

Theories about how deliberative democracy functions have a focus on individuals and communicative participation in society. Jürgen Habermas claimed that citizen participation and informed communication govern deliberation, ensuring that the outcomes in democratic processes are legitimate (Olson, 2011). John Rawls theorised about an individual's self-interest and how society should be fair and equal to all its members (Rawls, 1995). Habermas and Rawls laid a base for theories on a deliberative democratic process. The main features of a society that functions in a deliberative democratic way are shared practices and as a part of those practices, shared rules (Rawls, 1995). The assumption is that individuals commit to constitutive rules, meaning that both implicit and explicit rules that are formalized in some way are internalized by members of a society and followed voluntarily. This means that a deliberative democracy requires collective intelligibility and cooperation between citizens. Individuals need to understand and agree with the rules of society. In addition, participation should be based on free will and political freedom (Bächtiger, Dryzek, Mansbridge and Warren, 2018). Surveillance and scrutinization of individuals through information that holds everyone under a microscope impacts behaviour (Liang et al., 2018). A question that this paper aims to answer is how this scrutiny can affect the individual in the criminal justice system and what implications it has regarding a deliberative democratic process.

**Surveillance and deliberative democracy**

In 2004, the European Parliament stated that mass surveillance by governments cannot be justified and only targeted surveillance could be argued for. This means that only individual cases could be looked at when monitoring personal data, and a reason for this was needed. Suspicions of criminal behaviour were a perquisite for this kind of surveillance and a warrant and legal basis needed to be provided. However, this idea changed as European collective efforts were made to fight terrorism after 2006 (Maras, 2012). The Data Retention Directive by the European Parliament and European Council justified storing big data of individuals to surveil possible terrorist activity. What is common with mass surveillance is that individuals believe that only people who are suspects of crime are monitored to some extent (Maras, 2012). The idea of individuals participating in democratic processes and voluntarily following rules to meet a shared goal becomes problematic when citizens aren't informed about how they are being monitored. The idea that individuals have the free will to follow constituted rules becomes skewed when they are being monitored without their knowledge. It could therefore be argued that data surveillance for criminal control is going against the traditional ideal of deliberative democracy.

Surveillance of information on individuals that is hidden is problematic, yet the consequences of open record-keeping of encounters with the criminal justice system poses a new threat: system avoidance. In the United States, the extent of surveillance when it comes to crime has experienced a dramatic increase in the past decades. Records about criminal behaviour or contact with legal authorities are stored, and 47 million US citizens had a file with a criminal justice agency in 2014 (Brayne, 2014). This mapping of criminal information has led to individuals avoiding institutions that keep records in general. Citizens that know of their records being collected avoid even educational, medical, financial and labour market institutions, to avoid records being kept, and are less prone to engage in civic or religious institutions

(Brayne, 2014). This shows how surveillance and the criminal justice system can form a population that is detached from a democratic society. When looking at deliberative democracy, this kind of surveillance could be a threat to the commitment of citizens, and participation in democratic processes can suffer through system avoidance.

### Rules and social credit

John Rawls focused on how society works in a just way and how rules govern individuals. He distinguished between two types of rules, namely constitutive and regulative rules. Constitutive rules are set, either implicitly or explicitly, when regulative rules are the justifications of actions that are taken. Regulative rules therefore focus more on the moral obligation that individuals experience when following a constitutive rule. As an example, each chapter of criminal law is a constitutive rule in the legal framework of a community, yet when a person does not steal because it is the “wrong thing to do," they are following their internalized regulative rules (Rawls, 1995). Explicit regulative rules are not necessarily formalized, yet citizens follow them since they share a mutual understanding of how they work. In a deliberative democracy, individuals need to understand these rules and have the desire and free will to follow them for a shared purpose.

When it comes to extreme cases of surveillance through data, it can be argued that both constitutive and regulatory rule-following are being monitored to the extent that a person's free will is minimized. In the case of the Chinese Social Credit System, both public and private spheres of life are being monitored. Naturally, crime also plays a part in the model. Criminal activity is monitored and stored in a personal credit score (Liang et al., 2018). The aim is to reduce crime by punishing citizens through the reduction of their social score, which has implications on what they can and cannot do, for instance when it comes to traveling. In practice, if a person jaywalks or steals from a shop and gets caught, their credit suffers. However, the extent of control reaches behaviour that is not deemed criminal, yet just “bad” by the government. The state surveils individuals by gathering their data and provides rewards or punishments for correct behaviour (Liang et al., 2018). Therefore, the government is not only monitoring the explicitly stated legal constitutive rules but also explicit rules of good conduct, as defined by the government. When thinking about Rawls´ regulatory rules and the inner processes of rule-following, it could be argued that by monitoring the behavioural traits of individuals, the government is trying to also control the regulatory rules that steer an individual's behaviour.

### Conclusion

By critically examining surveillance and its impacts on behaviour in criminal justice systems, some implications for deliberative democratic processes are evident. Although surveillance might minimize criminal behaviour in society, it impacts behaviour on an individual level. It can be argued that the effects limit the agency and free will of citizens, hindering their participation in a deliberative democratic process.

Concealed surveillance of personal data raises questions about whether individuals are being voluntarily monitored, or whether it is a form of control, that threatens the idea of deliberative democracy, where participation in rule-following is voluntary. Moreover, individuals that are aware of the scrutiny of surveillance avoid systems that leave information behind completely. System avoidance creates a subpopulation of individuals who don't participate in democratic processes, which limits deliberation in those societies. Finally, in extreme cases such as social credit that is not merely based on criminal activity, but social interactions and behaviour, the implications of a marginalized group of avoidance could be far more problematic when thinking about participation in society and democracy.

## SRMs and Free Will within Deliberative Democracy: The Foundational Ideal of Respect explored – Nuura H. Naboulsi

In this section, we will argue that SRMs undermine the foundational ideal of free will within deliberative democracy by limiting the sphere for respect via computation. We do this by combining empirical examples of a computed human sphere with theories of free will and respect. This argument begins from the premise that human agency can alter historical conditions of possible experience, thus recognizing the continental tradition of thought (Rosen, 1999, 665).

The early stages of digitizing identity and reputation through various rating systems are evident in the Chinese social score and Western applications, such as Peeple, Yelp, Airbnb and DateCheck. They represent reputation applications, using behavioural economics, that fail to capture the full spectrum of human experience in computing it to a game that incentivizes strategic communication (Botsman 2017, 168-185) and aims to suppress collective expression (King et al., 2013). A Chinese person's score depends on their online friends' activities. This is highlighted by the system providers by warnings of befriending people of a lower score and exemplified in chat rooms where average score people seek high scorers (ibid.). Whether this is rooted in a psychological obsession of feeding a sense of control (Botsman, 2017, 168), an imposed emerging economic rationale (Zuboff, 2015), or something else is currently debated. Nonetheless, an intuitive anti-democratic impression of SRMs can be deduced with a practical application of certain conceptualized ideals attached to deliberative democracy, namely free will, and respect as a tangible part of it.

The notion of free will is highly contested and an intrinsic part of its discussion is on morality (Mele, 2014). In all theories of deliberation, the ideal of mutual respect is central (Bächtiger et al., 2018, 4). In practice, this refers to individuals actively listening and trying to understand the meaning of a speaker's statements, rather than perceiving those statements as objects to be dismissed, demeaned, manipulated or destroyed (ibid.). This ideal has been institutionalized to varying degrees globally, e.g. in the United Nation's Universal Declaration of Human Rights. It is beyond the scope of this paper to engage in an in-depth exploration of the various conceptualizations of mutual respect. Yet, for the purpose of the argument, it can be stated that in deliberative democracy the ideal of respect is widely agreed on, evident in its institutionalization and that violations against it are often controlled by different measures, including for example criminal law. Further, we can think of respect as a sphere, in which we are free to navigate, and which exists in direct and indirect interactions with others, who are also committed to the ideal.

The SRMs invade this sphere of respect by computing aspects of it. The Chinese Social Credit Score does this via the above-mentioned discouragement to interact with individuals of a lower score due to the fact that it can affect one's own ranking. In doing this, the statements of those individuals become "objects to be dismissed or demeaned". The personal score, of course, consists of several (some unknown) factors, and these are likely not neutral toward treating all individuals with equal respect. For instance, if one is unexpectedly unable to pay rent due to an illness that results in unemployment, that worsens one's economic situation and thus, drops one's score. This situation does not necessarily imply that other people lose their innate respect for this individual. Yet, the gamified sphere for respect within the score system encourages exactly that. In the interest of protecting one's own score, people are incentivized to abandon previous conceptions of mutual respect and to act according to ideals introduced by the score system. Therefore, the system invades and shapes the sphere for respect in ways that are contradictory to fundamental conceptions of respect in deliberative democracy. It further undermines free will, by imposing computed, institutionalized structures upon the social sphere, hitherto reserved for the pursuit of conversation oriented toward achieving mutual understanding and deliberation.  SRM controverts the above-mentioned ideals of communicative rationality which operate at the background of culture to make it a product of its subjects' interests. That is not to say there are no other invasive social-economic mechanisms in operation, nor that the social sphere is currently intact. We only observe that the described functions of SRMs undermine mutual respect, communicative rationality, and an ethos underlying deliberative democracy.

The SRM ideals reflect the interests of SRM providers, in the explored scenario of the Chinese government and business leaders. Those ideals might or might not align with deliberative conceptions of society. What can be established from this short exploration is that the above provided example based on the known functions of the Chinese score system violates certain ideals of deliberative models. It justifies and encourages direct discrimination based on at least economic income, which is affected by multidimensional socio-economic factors — including ones independent of individual capabilities, e.g. inequality (Wade, 2014, 322). This, in turn, is directly contradicts what the Universal Declaration of Human Rights, and the fundamental ideal of respect, tacitly recommend as the basis for deliberation.

The generation at the intersection of institutionalized deliberative ideals and its emerging SRM substitutes has the potential to encourage questioning the gradual, discrete institutional replacement of democratic ideals. We have attempted this by presenting a case, in which the contradiction of ideals can be illustrated in a clear theoretical manner. Speaking from this intersection, our argument cannot be divorced from the contextual conditions of its historical emergence, which heavily dictates our understanding of democracy among other commitments (Critchley, 2013, 11). Nevertheless, we have sought to formulate our argument on theoretical and empirical grounds without relying on an essentialist approach to the human sphere. We would like to highlight this by demonstrating how the argument built so far can be applied notwithstanding the position on free will, which is a highly contested topic (Mele, 2014). As ubiquitous as the notion of free will might be, the idea that certain spheres of human life have been kept relatively separated from political and economic interests so far might achieve consensus. They have been apart in the sense that the level of their impact on human will, however one might view it, has never reached the *intensity* of SRMs due to the scale of these applied novel technologies. Therefore, the fact that ideal of respect as a rather tangible concept that has been rethought repeatedly in human history, and (as a result) changed, ought to provoke the following question: Whether or not free will is illusionary, do we wish to let the SRM rationale enter more deeply into our private spheres of life in such a way that profoundly it shapes our capabilities in *all* spheres? This is one of the questions we propose should be addressed in further discussions on SRMs and computation more generally.

We have illustrated how the Chinese social score undermines deliberative democracy and free will by limiting the sphere for the foundational ideal of respect via methods of computation. It is beyond the scope of this paper to discuss all known SRMs, which might or might not function to similarly invade the human sphere, as some have suggested (Botsman, 2017). We have demonstrated how the computed respect within the Chinese model is *limiting* the domain of respect by encouraging discrimination. This is because the created scores are interdependent and consist of factors such as economic income, which, individually should not necessarily shape the sphere for respect between individuals. Yet the score encourages strategic communication, discrimination and imposes certain models and ideals. The explored case was shown to violate the ideals of deliberative democracy, free will, and respect. We have argued that the incentivized strategic communication and discrimination are grounded on political and economic interests, which deserve scrutiny in the face of their gradual institutionalization. We have further attempted to appeal to both sides in the debate on free will, by questioning whether it is desirable to allow these specific interests to gain a more prominent position in the human sphere via computation. It is the task of further research to engage in a more in-depth analysis of the limitations of computation, and the challenges it poses to deliberative democracy.

## Big Nudging – a Liberal Equivalent of an SRM? Implications for Autonomy, Free Will, and Moral Judgement – Johan Wahlsten

As the purpose of SRMs is to shape and control citizen behaviour towards more desirable patterns of conduct, they can also be examined as a subsequent step in the relatively novel conceptualization of *big nudging* (Helbing, 2019a, 28). Nudging is a notable policy trend of the last two decades, based on paternalism, which holds that governments and private actors ought to push or “nudge” individuals towards conduct that is more beneficial both for the individual and society. By altering the choice architecture of individuals nudges aim to incite favourable behaviour (see, e.g., Thaler and Sunstein 2008). Specifically, big nudging is, simply stated, the behavioural aspirations of shaping individual's behaviour enhanced immensely by the emergence of big data[1] (Helbing, 2019a, 38). Sætra posits that as big data allows for the gathering of significantly more information from individuals and their actions than previously, it makes nudging both highly more efficient and customisable, transforming nudge to a shove (Sætra, 2019, 2–3).[2] Yeung uses digital gerrymandering, i.e., voting manipulation, as an extreme illustration of such control (Yeung, 2017, 12).

The affiliations between technocratic big nudging and SRMs also apply to some extent to the Chinese SCS, even if at this moment the SCS does not intend to hide its paternalism and no high-level algorithmic governance is currently utilized in it (Creemers, 2018, 26–27). As Creemers has luminously pointed out, the SCS is a logical continuation of the Chinese political tradition that emphasizes the state's role as the supreme moral and behavioural authority. (ibid., 5–7). For Helbing, big nudging as a centralized top-to-bottom technocratic behavioural control is in many respects an equivalent of such a totalitarian regime, but "with a rosy cover" (Helbing et al., 2019, 80). In what follows, we will argue not only that the two are not dissimilar, but that forceful behavioural designs empowered by recent technological development undermine free will, autonomy, and moral judgement, and thus deliberative democracy, to a greater extent than the run-of-the-mill nudging. Plenty of time and space have been dedicated to debates on the conundrums and possibilities of nudging. Here, we do not have the opportunity to consider these debates in detail, but rather we will consider some relevant issues in the context of large-scale endeavours for behavioural modification made possible by the advancement of technology.

## Undermining the cornerstones of (deliberative) democracy

A relatively frequent argument against individualised social control empowered by big data and operating through a feedback loop is that it crams us to "filter bubbles", where diversified and surprising experiences, required for proper democratic participation (Yeung, 2017, 15, 17–18) and creative thinking might be discouraged or even abolished and social cohesion destabilized (Helbing et al 2019., 77, 80). What is more, as Helbing proposes, when behavioural goals are set externally, personal development is hindered and democratic pluralism thwarted (ibid., 85). The concerns

[1] Big Data is often characterized by Doug Laney's "three Vs": high volume, velocity, and variety of data (Laney, 2001). Based on a review of existing literature and definitions De Mauro et al. have offered a consensual definition: "*Big Data represents the Information assets characterized by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value.*" (De Mauro et al., 2015).

[2] For other discussions on big nudging see Helbing, 2019b (especially chapters 4, 7 and 11). Others call it "hypernudging" (Mills, 2019; Yeung, 2017). These scholars have somewhat different conceptualizations of the term, which reflects its novel status.

about data collection and informational privacy are also valid in these discussions (Yeung 2017, 9–10).

However, in our view, these are not necessarily the most crucial issues in big nudges and SRMs. As many scholars have noted, already the more regular examples of behaviour control by nudging, such as healthier snacks placed at eye level in cafeterias, can have negative effects on a citizen's "control over her own deliberation" and possibility to consider alternatives (Hausman and Welch, 2010, 130). Consequently, they can be understood to bypass the individual's conscious decision-making process (Mills, 2013, 32), and in essence then challenge intentional action (Searle, 1995, 6–7). More importantly, if the behavioural expectations and parameters are not transparent, these potent controls can also deprive individuals of an essential prerequisite for both political freedom and moral responsibility and thus also deliberative democracy; namely they challenge free will (Mele, 2014, 6). Such deceptions and smokescreens are probable because behavioural controls targeting psychological mechanisms are most effective when they are enacted "in the dark" (Bovens, 2008, 3). Such a possibility is also more likely if machine learning is heavily involved, due to the opaque black-box nature of algorithmic governance (Creemers, 2018, 27; Perel and Elkin-Koren, 2015, 482, 488).

As Sætra argues, behavioural guidance with low or non-existent transparency, targeting the pre-rational and subliminal functions of the individual by manipulating choice architecture, undermines both negative and positive liberty. How big nudges or credit scores meddle efficiently with the choice architecture of citizens can also impede with one's options, which according to Sætra, and the prominent philosopher Joseph Raz, makes them coercive (Sætra, 2019, 5–7; Raz 1986, 377–378). And the more effective and secretive this manipulation is, the more detrimental the effects are for liberty, autonomy and free will. If we agree with Sætra that big nudging is coercive, it is obvious that covert behavioural manipulation reinforced by technological developments can be also viewed as an outright challenge against the widely accepted ideal of the absence of coercive power in deliberative democracy (Bächtiger et al., 2018, 5).

Likewise, if we accept that covert and robust behavioural and social controls use *psychological force* as their means of coercion (Sætra, 2019, 7–8), then it is evident that what Alfred Mele calls *modest free will* – an individual having free will to the extent he is not subject to undue force – is under threat (Mele, 2014, 78). And as free will is often accepted to be an essential condition for moral responsibility, [3] strong behavioural controls may also put this facet of our humanity under contestation. Big nudges and SRMs are malignant to (collective) moral responsibility also in the context of freedom of action. Seamus Miller importantly emphasizes that any moral responsibility prerequisites an intentional action (Miller, 2010, 122). But if individuals are shoved towards certain behaviour unknowingly or interfered in their options with a menace of harm if a different option is chosen, then what can we say

[3] Some however neglect free will is a precondition for moral responsibility (see "Free Will", no date, paragraph 9–10).

about the existence of intentionality or autonomy, and therefore of moral responsibility? Who is responsible for actions done under forceful behavioural control? At the very least, big data powered behavioural control can weaken moral judgement, since internal conscious deliberations might be overridden. The importance of asking *why I ought to or not ought to do what I do* should not be underestimated.

### Away with manipulation, down with deliberation

It seems apparent then that the simultaneous developments in technology and behavioural policy aspirations have grave implications for our autonomy, free will, and democracy. Even if SRMs are yet to be developed as such in Europe or North America, we claim that big nudging is their close relative and they are not necessarily in accordance with liberal ideals of democracy. Some nudges may be benign (John et al., 2009, 366–367), but when technology increases its effectiveness and allows for more sophisticated methods of behavioural control, with possible elements of algorithmic governance and machine learning, the dangers grow increasingly. The urge for holistic top-to-bottom social and behavioural governance and circumscribing of liberty and self-government ought to be resisted. As Sætra, too, suggests, rational persuasion is much more preferable for this if we wish to preserve autonomy and liberty (Sætra, 2019, 7). In complex systems, short-cuts circumventing practical reasonable persuasion and debate for altering people's behaviour are easily ill-fated. There is no magic wand for complicated issues. Even when we accept bounded rationality, we should not settle for big nudging and other machine-enhanced means of advanced behavioural control, but have aspirations for better (Niemeyer, 2014, 31). If the issues mentioned are allowed to persist, the age of digital enlightenment and informed citizens will be a far cry, to paraphrase Immanuel Kant for the 21st century (Kant, 1784).

## Conclusion

In this paper, we have argued that:

- Surveillance is anti-democratic due to loss of privacy, freedom and autonomy and poses threats to democratic principles such as civic participation and plurality. More research, however, is needed before SRMs can be entirely ruled out as non-compliant with Western values.
- There are two ways in which SRMs threaten Habermas' notion of deliberative democracy from a privacy-perspective: lack of privacy in SRMs has a high potential to erode diversity among citizens by creating conformity pressures, and the black-box nature of the SRM can inhibit collective deliberation on informational self-determination. These issues challenge Habermas' reasoning advocating communicative rationality that is needed to debate common interests in the public sphere, which he identifies as foundational for deliberative democracy.

- The effects of surveillance limit the agency and free will of citizens and hinder their participation in a deliberative democratic process. Concealed surveillance of personal data raises questions about whether individuals are being voluntarily monitored, or whether it is a form of control that threatens the idea of deliberative democracy, where participation in norm-following is voluntary. In turn, individuals who are aware of the scrutiny of surveillance may completely avoid systems that might record and share information about them. This system avoidance then could exclude a group from democratic participation.
- The Chinese social score undermines the foundational ideal of free will within deliberative democracy by limiting the sphere for respect via computation. It further incentivizes strategic communication and discrimination, which are founded on political and economic interests. The model is directly against institutionalized ideals of deliberative democracy and respect and poses a challenge to their future.
- Big nudges and SRMs are both forceful behavioural designs empowered by recent technological developments. They undermine free will, autonomy, and moral judgment, and thus democracy to a greater extent than previous means of social control. If covert, this manipulation further can be viewed to challenge the absence of coercive power in deliberative democracy.

Through these five interconnected sections, we have attempted to shed light on the complex implications of SRMs on deliberative democracy. Deriving from theoretical and empirical accounts, we have found that SRMs have the potential to undermine different aspects of deliberative democracy. Throughout this paper, we have echoed that despite this development, the incorporation of more holistic behavioural control in democratic nations might become reality, evident in current surveillance and data manipulation practices.

## References

Bächtiger, A., Dryzek, J. S., Mansbridge, J., & Warren, M. E. (Eds.). (2018). *The Oxford handbook of deliberative democracy*. Oxford University Press, Oxford.

Botsman, R. (2017). *Who can you trust?: how technology brought us together–and why it could drive us apart*. Penguin UK, London.

Bovens, L. (2009). The ethics of nudge. In Till Grüne-Yanoff and S.O. Hansson (Eds.), *Preference change: Approaches from Philosophy, Economics and Psychology* (pp. 207-219). Springer, Dordrecht.

Brayne, S. (2014). Surveillance and System Avoidance. *American Sociological Review, 79*(3), 367-391. doi: 10.1177/0003122414530398

Campbell-Verduyn, M., Goguen, M. and Porter, T. (2017). Big Data and algorithmic governance: the case of financial practices. *New Political Economy*, 22(2), 219-236. doi: 10.1080/13563467.2016.1216533

Chen, Y. & Cheung, A. (2017). The Transparent Self Under Big Data Profiling: Privacy and Chinese Legislation on the Social Credit System. *The Journal of Comparative Law*, 12(2), 356-378. University of Hong Kong Faculty of Law Research Paper No. 2017/011
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2992537

Liang, F., Das, V., Kostyuk, N., & Hussain, M. M. (2018). Constructing a Data-Driven Society: China's Social Credit System as a State Surveillance Infrastructure. *Policy & Internet*, *10*(4), 415-453. doi: 10.1002/poi3.183

Olson, K. (2011). Deliberative democracy. In B. Fultner (Ed.), *Jürgen Habermas: Key*
*Concepts* (pp. 140-155). Acumen Publishing, Durham.
doi:10.1017/UPO9781844654741.008

Couldry, N. (2017). Surveillance-democracy. *Journal of Information Technology & Politics*, 14(2), 1-13. doi: 10.1080/19331681.2017.1309310

Creemers, R. (2018). China's Social Credit System: An Evolving Practice of Control. *Available at SSRN*: http://dx.doi.org/10.2139/ssrn.3175792.

Critchley, S. (2013). *Continental Philosophy: A Very Short Introduction*. Oxford University Press, Oxford.

De Mauro, A., Greco, M., & Grimaldi, M. (2015, February). What is big data? A consensual definition and a review of key research topics. In *AIP conference proceedings* (Vol. 1644, No. 1, pp. 97-104). AIP. doi:10.1063/1.4907823

Dorfman, A. (2014). Repression by any other name. *Guernica.* Retrieved from https://www.guernicamag.com/repression-by-any-other-name/

Floridi, L. (2014). *The Fourth Revolution: How the Infosphere is Reshaping Human Reality*. Oxford University Press, Oxford.

Free Will. In *Internet Encyclopedia of Philosophy*. Retrieved from https://www.iep.utm.edu/freewill/#H1

Fuchs, C. (2011). Towards an alternative concept of privacy. *Journal of Information, Communication and Ethics in Society*, *9*(4), 220-237.

Giroux, H. A. (2015) Totalitarian Paranoia in the Post-Orwellian Surveillance State. *Cultural Studies*, *29*(2), 108-140. doi: 10.1080/09502386.2014.917118

Habermas, J. (2003). Making sense of the EU: Toward a Cosmopolitan Europe. *Journal of Democracy*, *14*(4), 86-100.

Habermas, J. (1996). *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy*. Polity Press, Cambridge.

Hausman D.M., Welch B. (2010). Debate: To Nudge or Not to Nudge. *The Journal of Political Philosophy 18*(1), 123-136. doi: 10.1111/j.1467-9760.2009.00351.x

Helbing, D. (2019a). Machine Intelligence: Blessing or Curse? It Depends on Us! In D. Helbing (Ed.), *Towards Digital Enlightenment. Essays on the Dark and Light Sides of the Digital Revolution* (pp. 25-40). Springer, Cham.

Helbing, D. (2019b). *Towards Digital Enlightenment. Essays on the Dark and Light Sides of the Digital Revolution*. Springer, Cham.

Helbing, D., Frey, B.S., Gigerenzer, G., Hafen, E., Hagner, M., Hostetter Y., van den Hoven., J., Zicari R.V., Zwitter, A. (2019). Will Democracy Survivce Big Data and Artificial Intelligence? In D. Helbing (Ed.), *Towards Digital Enlightenment. Essays on the Dark and Light Sides of the Digital Revolution* (pp. 73–98). Springer, Cham.

Hill, C. (2017). How Can Liberal Democracies Defend Themselves against Tyranny? *Perspectives on Political Science*, *46*(4), 243-246. doi:10.1080/10457097.2017.1355136

John, P., Smith G., Stoker G. (2009). Nudge Nudge, Think Think: Two Strategies for Changing Civic Behaviour. *Political Quarterly, 80*(3), 361-370. doi: 10.1111/j.1467–923X.2009.02001.x

Kant, I. (1784). What is Enlightenment? Retrieved 9.1.2020. http://www.allmendeberlin.de/What-is-Enlightenment.pdf

King, G., Pan, J., & Roberts, M. E. (2013). How censorship in China allows government criticism but silences collective expression. *American Political Science Review*, *107*(2), 326-343.

Kostka, G., & Antoine, L. (2018). Fostering Model Citizenship: Behavioral Responses to China’s Emerging Social Credit Systems. *Available at SSRN*: http://dx.doi.org/10.2139/ssrn.3305724

Laney, D. (2001). 3D data management: controlling data volume, velocity and variety. M*ETA Group Research Note, 6*(70).

Macnish, K. (2014). Just surveillance? Towards a normative theory of surveillance. *Surveillance & Society*, *12*(1), 142-153.

Maras, M. (2012). The social consequences of a mass surveillance measure: What happens when we become the ‘others’? *International Journal Of Law, Crime And Justice*, *40*(2), 65-81. doi: 10.1016/j.ijlcj.2011.08.002

Mele, A.R. (2014). *Free. Why Science Hasn’t Disproved Free Will.* Oxford, Oxford University Press.

Miller, S. (2010). *The Moral Foundations of Social Institutions*. Cambridge, Cambridge University Press.

Mills, C. (2013). Why Nudges Matter: A Reply to Goodwin. *Politics, 33*(1), 28-36. doi: 10.1111/j.1467-9256.2012.01450.x

Mills, S. (2019). Into Hyperspace: An Analysis of Hypernudges and Personalised Behavioural Science. *Available at SSRN*: https://ssrn.com/abstract=3420211

Niemeyer, S. (2014). A Defence of (Deliberative) Democracy in the Anthropocene. *Ethical Perspectives, 21*(1), 15-45. doi: 10.2143/EP.21.1.3017285

Perel, M., Elkin-Koren, N. (2015). Accountability in Algorithmic Copyright Enforcement. *Stanford Technology Law Review*, *19*, 473-533.

Rawls, J. (1955). Two Concepts of Rules. *The Philosophical Review*, *64*(1), 3. doi:10.2307/2182230

Raz, J. (1986). *The morality of Freedom.* Oxford University Press, Oxford.

Rosen, M. (1999). Continental Philosophy from Hegel. In A. Grayling (Ed.), *Philosophy 2: Further through the Subject*. Oxford University Press, Oxford.

Sætra, H.S., (2019). When nudge comes to shove: Liberty and nudging in the era of big data. *Technology in Society 59*. https://doi.org/10.1016/j.techsoc.2019.04.006

Schwartz, P. (2000). Privacy and Democracy in Cyberspace. *Available at SSRN:* https://ssrn.com/abstract=205449

Searle, J.R. (1995). *The Construction of Social Reality*. The Free Press, New York.

Mac Sithigh, D., & Siems, M. (2019). The Chinese social credit system: A model for other countries?. *EUI Department of Law Research Paper*, (2019/01). https://cadmus.eui.eu/bitstream/handle/1814/60424/LAW_2019_01.pdf?sequence=1

Sunstein, C.R. (2003). *Why Societies Need Dissent*. Harvard University Press, Cambridge MA.

Susen, S. (2018). Jürgen Habermas: Between Democratic Deliberation and Deliberative Democracy. In Ruth Wodak and Bernhard Forchtner (Eds.), *The Routledge Handbook of Language and Politics*. Routledge, London.

Tene, O., Polonetsky, J. (2011-2012). Privacy in the Age of Big Data: Time for Big Decisions. *Stanford Law Review Online, 64,* 63-69.

Thaler, R., Sunstein R. (2008). *Nudge: improving decisions about health, wealth, and happiness*. Yale University Press, New Haven.

UN General Assembly (1948). Universal Declaration of Human Rights. *United Nations,* 217 (III) A, 1948, Paris.

Vedder, A., & Naudts, L. (2017). Accountability for the use of algorithms in a big data environment. *International Review of Law, Computers & Technology*, *31*(2), 206-224.

Wade, R. (2014). Growth, Inequality and Poverty: Evidence, Arguments, and Economists. In J. Ravenhill (Ed.), *Global Political Economy*. Oxford University Press, Oxford.

Westin, A. F. (1968). Privacy and freedom. *Washington and Lee Law Review*, *25*(1), 166.

Wiklund, H. (2005). A Habermasian analysis of the deliberative democratic potential of ICT-enabled services in Swedish municipalities. *New Media & Society*, 7(2), 247-270.

Yeung, K. (2017). 'Hypernudge': Big Data as a mode of regulation by design. *Information, Communication & Society*, *20*(1), 118-136.

Zuboff, S. (2015). Big other: surveillance capitalism and the prospects of an information civilization. *Journal of Information Technology*, *30*(1), 75-89. doi:10.1057/jit.2015.5

**SM Amadae** is a University Lecturer of Politics in the Faculty of Social Sciences at the University of Helsinki, and a 2020 Berggruen Fellow at the Center for the Advanced Study of the Social and Behavioral Sciences, Stanford University. She is also affiliated with Science and Technology Studies, Massachusetts Institute of Technology; and the Centre for the Study of Existential Risk, University of Cambridge. Amadae's books include *Prisoners of Reason: Game Theory and Neoliberal Political Economy* (Cambridge University Press 2016); and *Rationalizing Capitalist Democracy: Cold War Origins of Rational Choice Liberalism* (University of Chicago Press 2003). Articles relevant to the *Computational Transformation of the Public Sphere* are: "Game Theory, Cheap Talk and Post-Truth Politics: David Lewis vs. John Searle on Reasons for Truth-Telling," *Journal for the Theory of Social Behavior*, Vol. 48:3, 2018; "Computable Rationality, NUTS, and the Nuclear Leviathan," chapter 6, in *The Decisionist Imagination: Democracy, Sovereignty and Social Science in the 20th Century*," ed. by Daniel Bessner and Nicolas Guilhot New York: Bergahn Books, 2018; with Shahar Avin, "Autonomy and Machine Learning as Risk Factors at the Interface of Nuclear Weapons, Computers and People," *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk: Euro-Atlantic Perspectives*, SIPRI, 2019, chapter 13, 105-118; and most recently, "Life as Algorithm," *Twenty-First Century Approaches to Literature: Futures*, ed. by Jenny Andersson and Sarah Kemp, Oxford University Press, forthcoming, 2020.

**MA Program, Global Politics and Communication** **amadae.com**