Universiteit Leiden
The Netherlands

Automated Influence and Social Agency: The Limitations of Regulatory State Power in Addressing Concerns Beyond Privacy

Zewen, Claire

Citation
Zewen, C. (2023). *Automated Influence and Social Agency: The Limitations of Regulatory State Power in Addressing Concerns Beyond Privacy*.

Version:	Not Applicable (or Unknown)
License:	License to inclusion and publication of a Bachelor or Master Thesis, 2023
Downloaded from:	https://hdl.handle.net/1887/3655588

Note: To cite this publication please use the final published version (if applicable).

Claire ZEWEN 24th August 2023

Automated Influence and Social Agency:
The Limitations of Regulatory State Power in Addressing Concerns Beyond Privacy

MA Thesis (20EC) by Claire ZEWEN

MA Philosophical Perspectives on Politics and the Economy
Supervisor: Dr. Dorota Mokrosinska

Universiteit
Leiden
The Netherlands

Abstract

In recent years, Automated Influence, understood as "the use of artificial intelligence to collect, integrate and analyse people's data, and to deliver targeted interventions based on this analysis, intended to shape their behaviour" (familiarly referred to as 'algorithms') has stirred up many debates among the public, as well as within academia (Benn & Lazar 2022, 127). While much of the discussion has focused primarily on issues of privacy in the light of Big Data, this thesis seeks to analyze how Automated Influence impacts the deliberative, discursive, and fundamentally social space on which society depends on, in particular for collective decision-making/politics. I argue that Automated Influence deployed on social media platforms violates people's fundamental interest in social agency, which is defined as the ability of a person to act and reflect on her own motives all the while taking part in the fundamentally social process of forming, defending, and adapting the reasons according to which she acts. Moreover, it undermines people's autonomy and social trust, which both serve as preconditions for their exercise social agency. After reviewing contemporary EU regulation seeking to address some of the problematic aspects related to Automated Influence, I explain why there cannot be a purely top-down approach to mitigating the harms emanating from Automated Influence, which results in my conclusion that only through educating people about its potential harms could mitigate the problem in the long run.

I. Introduction

In her seminal book *The Age of Surveillance Capitalism: The Fight for a Human Future at the Frontier of New Power*, Shoshana Zuboff delivers an impressive critique of contemporary mass surveillance practices underpinned by the logic of capitalism. Reaching beyond the traditional paradigm of surveillance as perpetrated by government agencies and involving the singling out of people of particular interest, Zuboff explores the emerging business model of converting behavioral data into powerful predictions about human behavior that are traded within an expanding network of surveillance capitalist industries. If Big Data, technically understood as the increase in quantity, scope, and velocity of data generation and exchange, serves as the underlying condition and expression of surveillance capitalism (Zuboff 2015, 77), Automated Influence reveals its new arm of power.

In a nutshell, Automated Influence can be understood as "the use of artificial intelligence to collect, integrate and analyse people's data, and to deliver targeted interventions based on this analysis, intended to shape their behaviour" toward profitable ends (Benn & Lazar 2022, 127). In other words, Automated Influence processes seek to influence behavior on the level of psychographically contrived groups, meaning on the level of an aggregation of people grouped together by algorithmic pattern detection. Automated Influence is familiarly equated with one of its technical instruments: algorithms. Both academic and layman debates about the potential harms resulting from "algorithms" has abounded in recent years, not least because of the increasing relevance of online social media platforms (SMP), which serve as the main hub for the types of Automated Influence of particular interest to this thesis' critique. Most of the discussions surrounding Automated Influence have problematized the issue of personal privacy, mass surveillance, the proliferation of misinformation, and their impact on democratic societies in particular.

Zuboff problematizes the creation of a "hive mind" as a result of Automated Influence, which she views primarily in terms of "instrumentarian power" (Zuboff 2019). The creation of a hive mind designates the process describing the "social confluence, in which group pressure and computational certainty replace politics and democracy, extinguishing the felt reality and social function of an individualized

existence" (ibid.). While Zuboff focuses her critique on what is lost as a result of replacing the uncertainty of individual, autonomous action with nudging techniques coordinating group behavior, this thesis' critique of Automated Influence problematizes its socially *divisive* impact. Focusing on the aggregate, rather than the individual level in analyzing the impact of Automated Influence, I will answer the question as to how the impact of Automated Influence can be gauged against the backdrop of people's fundamental interest in social agency and autonomy, and how it affects social trust.

More precisely, what this thesis sets out to do is to ground the concern with regard to Automated Influence and the practices of actors and technologies involved in it in an ethical framework of fundamental interests. In particular, it will be argued that Automated Influence has problematic implications for the respect of fundamental interests. Fundamental interests concern the conditions necessary for the existence of personal attributes considered universally and non-contingently valuable. In this regards, the main object of concern for this thesis is social agency, which is necessarily conditioned on autonomy, conceived from an *inter*- rather than an *intra*personal approach and which can only flourish in conditions of a relatively stable level of social trust.

Thus, to reformulate the guiding question, this thesis sets out to inquire how Automated Influence impacts autonomy and social trust, which are identified as conditions for social agency. Unlike many previous critiques of Automated Influence, this thesis sets out to examine issues beyond personal privacy, which traditionally confines the issue to the sensitivity of information it allows to gather on people through data analytics.

In the first chapter (second section), I will lay bare the meaning and ethical weight of fundamental interests as opposed to basic needs by leaning on Fabian Schuppert's (2013) distinction. Specifically, I will explain the main objects of fundamental interests, namely social agency. I argue that autonomy, conceived as a fundamentally social, interpersonal phenomenon, as well as social trust, are necessary for social agency. Having specified and explained the ethical objects of interest framing the discussion, I will move on to the second chapter (third section) in which I provide a

rather technical, more in-depth explanation of what Automated Influence is and what kind of systems it comprises. Particular attention will be devoted to recommender systems, the algorithmic structures governing content recommendation on SMP. Moreover, I will discuss micro-targeted advertising as another relevant system of Automated Influence, exemplified by the Cambridge Analytica-Facebook scandal. While the latter is rather straightforward in its ethical wrongdoings, the implications of recommender systems are more difficult to discern. The third chapter (fourth section) will be dedicated to exploring the harm inflicted by Automated Influence, including recommender systems. I will first focus on the conventional concerns associated with Automated Influence processes, which I identify as privacy, consent, and information sensitivity. Then, I shift my focus to interpersonal autonomy, with a particular interest in how Automated Influence can affect the motives and reasons upon which we act. The main observation is that Automated Influence interferes with the social process of grounding the normativity of reasons. Moreover, it impacts social relations in terms of undermining social trust, as people become more and more skeptical with regards to the autonomy of others. The fourth chapter (fifth section) is dedicated to an exploration of existing political responses to the perceived threats of Big Data and its processing through artificial intelligence. In particular, I will identify the achievements, as well as the shortcomings of the European Union's General Data Protection Regulation and its Digital Services Act.[1] Based on these insights, I conclude the thesis by pointing to the need for an alternative approach to addressing the concerns of Automated Influence. As I argue, it is only by complementing top-down measures by bottom-up initiatives to educate people about the potential risks associated with Automated Influence that the problem can be mitigated in the long run.

II. Fundamental Interests, Social Agency, and Autonomy

II.1. Social Agency as a Fundamental Interest

In order to properly frame the discussion regarding the ethical wrongs of Automated Influence, it is necessary to explain what this thesis' object of concern is and what kind of moral status it occupies. The focus will be on the social, intersubjective space

[1] The Digital Services Act has not yet entered into force. Thus, its "achievements" are gauged in a prospective sense based on what is set to be included in the piece of legislation.

for dialogue and action and the way in which Automated Influence deployed on SMP impacts the wellbeing of thereof.

In reviewing the most prominent accounts of justice, Fabian Schuppert (2013) highlights the importance of complementing the often invoked concept of basic needs with the idea of "fundamental interests". Basic needs are "needs which possess absolute necessity for the existence of the needing being" and which she cannot be or do without (ibid., 29). They are the means necessary to achieve "valuable non-contingent end(s)" such as life, harm-avoidance, and agency (in a minimal sense)(ibid.). These ends are non-contingent, i.e., independent of any variable to the person they concern, insofar as they imply a claim of existential necessity to the person in question. The more strictly and narrowly the values/ends are conceived, the more likely they can legitimately be objects of basic-need claims. However, given the unique moral fortitude of basic-need claims, which lies in its appeal to existential, non-contingent requirements, it cannot encapsulate the totality of ethically weighty concerns. As Schuppert (2013) argues, some values, which are not of absolute necessity to the existence of the person in the sense of her survival, but are still of particular normative weight, can be conceptualized as "fundamental interests."

The object(s) of fundamental interests refer(s) to the "group of non-contingent needs which are of *necessary instrumental value for achieving the universally valuable ends of persons*" (ibid., 32; emphasis added). These needs are non-contingent – or universal – insofar as they enable people to be agents and agency is of fundamental value to all persons and therefore constitutes a fundamental interest. The non-contingency of the need for agency derives from the fact that it is a precondition for the freedom (as opposed to the domination) of individuals. In particular, they are of "necessary instrumental value for" individuals' freedom to act. Thus, according to Schuppert and in congruence with this thesis' focus, the "universally valuable ends of persons" designate, first and foremost, the freedom to act on the basis of one's own intentions, desires, and beliefs. Universally valuable ends are shared, non-contingent, and – as a result – bear fundamental-interest claims (ibid., 34). They designate the conditions enabling persons to be *free and rational agents*.

Further above, I pointed out that agency is a basic need rather than a fundamental interest. This is the case for minimal agency, i.e., the *capacity* of the individual to

form and act on the basis of intentions and identify, process, and incorporate reasons into decisions concerning her actions (ibid., 30). It is a legitimate object of a basic-need claim since, without it, an individual becomes a passive patient instead of an agent, which is an intrinsic, essential element to personhood (ibid.).

The question whether agency fulfills the criteria of basic-need claims depends on how narrowly or widely one conceives of agency and its functions and requirements. While minimal agency is *necessary* "for being able to act in the light and on the basis of one's intentions and reasons", Schuppert argues that it does not necessarily guarantee that a person act freely and autonomously (ibid.). It merely designates her capacity, i.e., her potential, to do so.

A thicker account of agency considers the conditions enabling autonomous reasoning and the freedom to act based on this reasoning. Such an account of agency cannot be object of a basic-need claim, since a lack thereof does not unequivocally threaten the existence of the person in question, nor her status as an agent. This points to the fact that basic-need claims – while potent in invoking a sense of moral urgency to their fulfillment, are rather limited in scope (ibid., 32). As such, they cannot account for a more substantive conception of agency that carefully considers the condition of autonomy and freedom. This is why Schuppert argues in favor of a social conception of agency, i.e., "social agency," as an object to a fundamental-interest claim.

II.1.a. Conditions for social agency

Every person has a fundamental interest in being a free and rational agent, regardless of their own particular conceptions of the good (ibid., 35). The conditions enabling them to be so extend over a range of domains, including the person's own physiological and psychological predispositions, but also the social, political, and economic context she finds herself in (ibid., 35). In other words, the social relations which the person finds herself involved in also play into her actual *ability* to exercise her social agency. The kind of conditions on the basis of which a person can be said to exercise her free, rational agency (i.e., her *social agency*) concern values of collective/social rather than exclusively individual interest, such as freedom, social equality, mutual inter-subjective recognition, and democratic fairness (ibid., 36). If the social relations the person finds herself and takes part in reflect these values, they enable her to exercise her social agency.

Thus, *social* agency highlights an intersubjective approach to autonomy and the freedom to act, meaning it emphasizes their nature as social states rather than natural conditions (ibid.). The kind of free, rational agency that can claim the status of social agency presupposes a social space of deliberation in which people exchange reasons on the basis of which they act and can be held accountable for. It is only within this intersubjective sphere that the reasons guiding people's actions can be gauged in terms of validity and authority. As Schuppert puts it himself:

> [W]e do not reason and act freely isolated from other beings but in a *shared social space, a normative and discursive space* in which we give, receive, defend and debate reasons for and against certain actions. While it is undeniably clear that we live in a social world, our account of rational agency claims that *being socially and intersubjectively connected is necessary for and constitutive of rational agency*, and not – as often suggested – an obstruction to it. In short, in order to be free rational agents we need to stand in certain relations to the other members of society, namely, in *relationships of mutual intersubjective recognition* (Pippin 2008). […] Phrased differently, free rational agency requires as a relation to others and as the basis for proper self-relation mutual, reciprocal recognition and discursive practices in which the agent can express her normative commitments and take responsibility for her actions and judgments. (Schuppert 2013, 36; emphasis added)

Thus, in addition to the fundamentally social "normative and discursive" space being a precondition for social agency, the latter is also contingent on the kind of relations by which the people in it are bound together (however loosely that may be). These relations are characterized by "mutual intersubjective recognition" of one another as having an equal claim to and capacity for social agency. This point also serves to underline the importance of relatively high and stable level of social trust for social agency. For the "mutual intersubjective recognition" of one another as equals, people have to be able to reasonably trust that their fellow members of society are equally able to exercise their social agency. Only through social relations of that nature can the "shared social space" play the functional role of grounding the normativity of reasons.

The giving, exchanging, adapting, adopting, challenging, and rejecting of reasons and motives does not always take explicit form. In other words, obviously not all actions and the motives/desires they are based on are debated between interlocutors in an act of deliberation. Rather, norms/standards of intelligibility are imbued in our common

horizon of meaning, the discursive tissue that holds together communities who share a language and, in a minimal sense, a culture.

Social agency and does not limit itself to a rationality primarily defined in terms of self-interest. Rather, it focuses on acting according to reasons that are not only the agent's own, but also comprehensible by others (i.e., "universally valuable ends") (ibid., 36). In other words, a person is said to exercise her social agency if she acts on the basis of reasons that hold some universal valuability. The shared reasons/ends are of intersubjective, rather than subjective nature, not necessarily in the sense that the exact motives guiding a social agent's action are endorsed by others, but rather that they are intelligible and, in some sense, valuable. Phrased differently, even if other persons which a social agent finds herself bound in social relations with do not identify with her actions and motives, the fact that they could understand the reasons underpinning the motives of the person makes the action following the motive a legitimate exercise of social agency.

II.2. Personal Autonomy

Thus far, we have established that social agency is a fundamental interest and that it is exercised in the intersubjective, social space. While Schuppert emphasizes the role of reasoning and rationality with regards to social agency, he foregoes a more in-depth, explicit discussion of personal autonomy. Personal autonomy precedes every argument about agency, as it designates the authority of the person to govern herself, derived from the fact "she alone can initiate her actions" (Buss & Westlund 2018). Thus, in addition to the aforementioned social conditions for social agency, i.e., freedom, social equality, and mutual inter-subjective recognition (Schuppert 2013, 36), first and foremost, a person has to be autonomous. Insofar as personal autonomy is a precondition for (social) agency, it can be said to be of fundamental-interest status.

It goes without saying that autonomy is a complex concept, one that varies in scope and substance within scholarly debate. Simply stated, the question whether or not a person is autonomous is determined by the extent to which she has power over the motives guiding her actions. Thus, while in a narrow sense, the agent will always herself be the source of her own actions, this does not imply that she is in control over

the motives guiding them (Buss & Westlund 2018). What, then, makes her *not* be in control of her own motives? In other words, at what point can external forces be said to have sufficiently overtaken the self-governing process for someone's autonomy to be impaired?

> Intuitively, an agent can fall under the sway of desires, or urges, or compulsions whose power is at odds with her own power as an agent; she can be moved by such impulses "in spite of herself." But in what sense, exactly, are such motives "external" to the agent herself? How can *their* power to move her fail to be a manifestation of *her* power to act? (Buss & Westlund 2018)

Naturally, this depends on how one substantially conceives of the self-governing process(es). This question has been subject to numerous philosophical debates. So-called "coherentists", for example, argue that whether an agent can be said to act autonomously depends on how the motive according to which she acts relates to her other desires and preferences, whether they are *coherent* with a broader set of volitions, notably her highest-order desires which represent her long-term plans and values (ibid.). "Responsive-to-reasoning" accounts of autonomy agree with the notion of a reflective, intrapersonal process conferring authority to some reasons rather than others. But rather than emphasizing the pre-eminence of highest-order preferences themselves, such accounts highlight the role of practical reasoning in supporting these preferences (ibid.).

II.3. From an intra- to an interpersonal approach

While there is merit to approaches to personal autonomy highlighting the reflective processes internal to the person, as Joel Anderson (2003) argues, they end up failing to draw a distinctive line between a 'change of heart' and self-betrayal. In other words, such conceptions of autonomy do not adequately address the question whether acting according to reasons incompatible with previously held reasons merely represents a change in preferences (or a failure to reason in favor of them), or rather constitutes an act against oneself, revealing a lack of autonomy.

This gap highlights the shaky grounds, which any conception of personal autonomy relying on an intrapersonal account of normative authority of preferences/reasons is based on. The failure to determine the grounds of the authority of a given desire leads Anderson to approach the question from an *inter*personal, social constructivist point

of view. While he does not reject the idea, most prominently developed by Harry Frankfurt, of an internal normative order to our desires, he insists that the "normative grip" of reasons always ultimately derives from socially given norms. To be clear, these norms can be challenged, or even rejected altogether. The point is, though, that even the challenging or rejection thereof takes place within the public social sphere given that it presupposes the recognition of the existence of those norms. In other words, even in challenging or rejecting norms, one still implicitly recognizes their existence.

This is not to say, however, that Anderson's conception of autonomy is underpinned by a teleological, i.e., a outcome-oriented ethics. The point is not that societal norms ought to tell us how to act, or even that they are always right in conferring normativity onto some motives and reasons rather than others. Rather, what Anderson refers to when making reference to norms are "standards of intelligibility" (ibid., 99). In other words, the normative grip of reasons according to which one acts derives from their generality in terms of their social recognition and validity, which are, in turn, conditioned on the wider community, however loosely conceived. This is in line with the distinction I drew further above between (minimal) agency and social agency and the "universally valuable ends" on which it is predicated.

To clarify the issue, I will borrow Charles Taylor's example (evoked by Anderson) of the desire of wiggling one's toes in the warm mud (Anderson 2003, 100). Intuitively, if a person experiencing this desire were to change her mind and refuse to act on it, no one would claim that she is, in fact, engaging in an act of self-betrayal. Nothing about the desire to wiggle her toes in the warm mud could reveal any reasons of essential relevance to her practical identity. Therefore, if she decided to not act according to it, hardly anyone would qualify it as an act against herself. The desire to wiggle one's toes in the mud simply does not have the kind of normative grip that others would have. In contrast, even a desire as seemingly insignificant as learning to play the piano could very well have more normative weight, depending on the reasons attached to it. For instance, one might find that practicing a musical instrument contributes to one's flourishing by developing a different skill and appealing to an aesthetic appreciation of art. Intuitively, it is not difficult to see how such a motive could be more deserving of a stronger claim to one's agency/freedom.

What these examples reveal is that the authority of commitments has an essentially and inevitably social dimension (ibid., 91). In other words, it is "norms of intelligibility" that determine whether a deviation from previously held commitments represents self-betrayal or simply a change of heart (ibid., 103). Such norms of intelligibility are defined by the social space in which agents can give and justify reasons. They are "embedded in and presupposed by social practices to which we are almost unavoidably committed" (ibid., 102). This does not mean that we, or the authenticity and authority of our commitments, are passive products of social norms. What it means is that we are always already acting and reflecting within a social space that influences us and, at the same time, is influenced *by* us. In Andersons' own words, the normative grip given by social norms of intelligibility "can be challenged, but it cannot be wished away by individuals" (ibid.).

The relational argument is that, without a public space in which members of society together, through language and other discursive practices, express, establish, and (re)evaluate norms of intelligibility and social values, there cannot be a self-governing individual. The self-governing process, determined by our capacity to reason autonomously, does not and – according to relational theories of agency and autonomy – *cannot* play itself out in isolation of the discursive, social field constructed over generations and extending over a given social community. There would simply not be the kind of reasons we have in mind when talking about social agency without a social community. It is in the pursuit of the latter and, especially, the numerous benefits of cooperation and interrelation, that certain values have been formed and certain reasons can find their social legitimation. In short, social norms arise – at least in part – from the requirements of living together in however loose of a community.

To conclude this chapter, social agency is exercised and conditioned upon a shared social space, which serves to ground the normativity of certain motives, reasons, and subsequent actions. This is generally an implicit process (i.e., discourse) but it can also be made explicit through deliberation. The social relations by which members of the shared space are bound together are mutual insofar as people recognize one another as equals in terms of their right to social agency and their capacity for

autonomy. Just like social agency fundamentally relies on intersubjectivity, so does autonomy. In other words, to reason autonomously does not mean to reason independently or in isolation of others. Quite to the contrary, one's autonomy is measured by the degree of universal intelligibility of the reasons and motives guiding one's action. The kind of autonomy that is a necessary condition for social agency is viewed from an interpersonal approach, emphasizing norms of intelligibility as grounding the normative force of reasons. This further qualification to autonomy not only serves to answer the question regarding when a change in motives can be considered as acting against oneself/self-betrayal rather than a simple change of heart. It also serves to resolve the conundrum regarding the internality or externality of motives and reasons. While many accounts of autonomy rely on the distinction of motives as internal to the agent in order to qualify the latter as autonomous, the interpersonal approach clarifies how this is the wrong distinction to make in the first place. The agent's desires are always formed and validated both internally and externally.

III. Automated Influence Explained

Having defined the ethical objects of interest in relation to which I will gauge the impact of Automated Influence, the next step consists in explaining what exactly is meant by the latter. Automated Influence refers to "the use of artificial intelligence to collect, integrate and analyse people's data, and to deliver targeted interventions based on this analysis, intended to shape their behaviour" toward profitable ends (Benn & Lazar 2022, 127). In other words, the underlying purpose of Automated Influence techniques, such as recommender algorithms deployed on SMP, is to influence people's attitudes and actions. The desire to influence people's behavior and worldviews is nothing new in itself; it has existed as long as the interest in power has. In other words, social, political and economic actors have always had a vested interest in shaping the behavior and attitudes of society, the electorate, and consumers. From marketing to political advertising, long before the existence of Automated Influence and SMP, social and political actors in both the private and public realm were willing to pay a lot in exchange for influence. Yet, the breakthroughs in data science and the novel scale of available data, instruments of data analytics (e.g., artificial intelligence), and the possibility to reach millions of people through one common

interface (SMP) has truly amounted to a revolution in the realm of communication, social outreach and influence. The scale, depth, reach and speed of both - relevant data supplies and influencing methods, the two integral elements to Automated Influence, have widened the horizon of possibilities in this regard.

The term 'Automated Influence' generally refers to systems composed of multiple processes serving to carry out its purpose, i.e., to influence people's beliefs and behavior – their attitudes and actions. Automated Influence mainly works by *targeting* users with specific content or ads through recommender algorithms based on insights gained from the analysis of data pertaining to the user. It also allows to *tailor* messages to particular users, meaning it can modify the messaging based on the inferred susceptibility of the user to different methods of persuasion, although tailoring is less commonly used than targeting (Benn & Lazar 2022, 127).

> These interventions aim to shape the user's behaviour—that is, they aim to raise the probability they will ultimately take some particular course of action—in order to realise some goal. Behaviour is, minimally, a function of one's beliefs and desires given one's option set. Automated Influence can shape each element. Search and newsfeed algorithms shape what we believe; ads and recommender systems prompt and direct our desires; platforms make some options available and attractive, while hiding others. (ibid., 127-28)

Automated Influence provides technologically new models for social outreach through digital media. It works on the basis of valuable insights into people's internal, psychological pathways by means of automated processes attributable to artificial intelligence (AI). AI provides insight into how a message is most pertinently conveyed to a specific users, at which intervals they have to be 'touched' - and how often, in order for a mental process to trigger a certain path of action desired by those who control the particular parameters of the mission carried out by Automated-Influence means. Thus, the psychological insights which Automated Influence relies on are concerned with measuring the current mental state of the user to be touched by the message (in order to gauge how to most efficiently influence them), as well as with engineering (or triggering) a mental state conducive to the desired course of action (Benn & Lazar 2022).

Recommender systems (a type of Automated Influence technique) feed on training data, which contains personal information we advertently provide SMP with (e.g.,

age, location, educational background, employment), as well as behavioral data (i.e., data inadvertently generated by users' interactions with the platform, e.g., what content attracts their attention for the longest, or how much time they spend browsing through their feed). By collecting (training) data at large scale and by means of automated data analysis through AI, Automated Influence systems can make inferences about the preferences and beliefs of platform users and study their emotional triggers and pathways of action (ibid.). SMP use recommender systems, familiarly referred to as algorithms, to target users with content that will keep them engaged (i.e., using the SMP's interface) for as long as possible. Social media platform firms' (SMPF) bottom line depends on increasing user engagement in order to grow revenues derived from third-party advertisers using the platform to advertise their products or services. The higher the overall user engagement on the SMP, the higher the prices SMPF can charge advertisers and the more revenue they will generate.

In addition to the use of Automated Influence techniques on SMP by SMPF, the unique service provided by the combination of these technologies can be contracted by third parties for any influencing campaign they wish to deploy. Potential objectives could vary from advertising to political campaigning. The Facebook-Cambridge-Analytica scandal unveiled in 2018 represents one of the more egregious uses of Automated Influence uncovered in recent years and falls under the category of a political influencing project. It deserves our particular attention for several reasons: first, it is an exemplary case of Automated Influence insofar as both targeting and tailoring techniques were utilized. In addition, the way that the relevant data was procured was particularly outrageous as people were not asked for consent about the use and processing of their information. Finally, the case reveals how much risk there is to the commercialization of data science, as the digital online field is now conceived as a playground for social engineering.

III.1. The Facebook-Cambridge-Analytica scandal

In 2018, former employee Christopher Wiley contacted the British newspaper *The Guardian* to blow the whistle on what he considered morally wrongful practices of the data-science company Cambridge Analytica (CA). The latter was created in 2014 to offer services to business and political actors promising to "change audience

behavior" (Osborne 2018). By the time of CA's inception, CEO Alexander Nix had worked in finance, strategic communication, data mining and analysis. He was president of SCL Group (Strategic Communication Laboratories), CA's holding company, itself active in the data-mining and analysis industry. Nix teamed up with Donald Trump's then chief executive campaign officer, Steve Bannon, who would come to be appointed as Trump's chief strategist and senior counselor, as well as American billionaire entrepreneur Robert Mercer to create CA. The name of the company was supposed to give it an academic façade purely by association of names, which was particularly to the liking of Steve Bannon, who saw himself and his ambitions through an intellectual veil, as Wiley later revealed to *The Guardian* (Cadwalladr & Graham-Harrison 2018).

Bannon was on a mission to mobilize a political messaging campaign based on insights from behavioral science combined with an arsenal of data and instruments of Automated Influence, all of which Cambridge Analytica – in collaboration with data scientists Aleksandr Kogan (at the time, lecturer at Cambridge University) and his "close working relationship" with Facebook could provide (ibid.). Bannon's mission was grounded in his core idea that to change politics, you need to be able to influence culture. In order to change culture "you have to first understand what the *units* of culture are. People are the units of culture. So if you want to change politics, you first have to change people (…)" (ibid.). Thus, in a nutshell, Steve Bannon's vision and Cambridge Analytica's mission consisted in targeting individuals and tailoring the intended messaging to their individual psychological profiles. The data was sourced through third-party apps installed on Facebook and the influencing of each individual was engineered through content creation, tailoring, and targeting.

Kogan had developed a quiz-style personality test app accessible on Facebook and even advertised on a crowdsourcing platform, where people were paid $1 to $2 to take the personality test. The results of these tests would serve as training data for the algorithm developed to target people. To take the test, users had to give the app access to their Facebook profiles and the ones of their networks of friends. This was a cheap and easy solution to harvest data from large swathes of people without having to reach them directly through the app (cite). Specifically, 'only' a few hundred thousand people had to take the quiz for CA to extract data from 87 million people (Hern

2018). On the basis of the information obtained from the data, the algorithms would construct the psychological profiles of those millions of people in order to not only target them with political content Cambridge Analytica, but also to tailor the messaging in the content according to the inferentially most effective way of persuading the person – in terms of their psychological profile. Consider the example of pro-jobs political messaging. Certainly, it is an uncontroversial objective to campaign for, but in adapting how exactly this message is relayed to people of different political leanings and personality traits makes a difference in its effectiveness.

> What that means in practice is that the same blandishment can be dressed up in different language for different personalities, creating the impression of a candidate who connects with voters on an emotional level. "If you're talking to a conscientious person" – one who ranks highly on the C part of the Ocean model – "you talk about the opportunity to succeed and the responsibility that a job gives you. If it's an open person, you talk about the opportunity to grow as a person. Talk to a neurotic person, and you emphasise the security that it gives to my family." (Hern 2018)

Thus, CA could target individuals with (political) ads that were tailored according to their inferred susceptibility to the message. While it is improbable that, on an individual scale, CA's influence campaign was effective in convincing liberal and left-leaning voters to vote for Trump, their targeting methods allowed them to specifically identify potential swing voters and non-voters in order to more efficiently deploy their influencing resources (Cadwallard & Graham-Harrison 2018). People with these predispositions would be more susceptible to the anti-establishment messaging the campaign often employed. And, as Benn and Lazar (2022, 142) write, when it comes to mobilizing political support in the context of elections, as well as broadly in terms of particular policy proposals, "the ability to sway a given group by a few percentage points, even a few fractions of a percentage point, can ultimately prove decisive (Heilman 2020)."

Eventually, Trump's presidential bid was a success. Research suggests that much of the campaign's success lay in gathering support of swing-voters (meaning voters previously not registered to vote Republican), as well as of people who usually did not vote at all (Hill, Hopkins & Huber 2021). While it will never be fully known to what extent CA's work helped in putting Trump in the White House, the kind of

techniques employed within the framework of Automated Influence could have played into the success of the campaign. Regardless of how effective CA was in influencing people to vote for Trump, the kind of question this thesis is seeking to answer is a more fundamental one: whether Automated Influence works to undermine social agency and autonomy of those who are affected by it. In other words, this thesis is concerned with whether Automated Influence impedes people's fundamental interest in social agency, as well as how it affects autonomy and social trust, which social agency is conditioned on.

IV. Automated Influence and the Threat to Autonomy and Social Agency

In the first chapter, I emphasized the importance of social relations of mutual recognition and the discursive space in which shared meanings and exchanging reasons represent the condition for personal autonomy and, thus, social agency. The following chapter is dedicated to evaluating whether and to what extent Automated Influence can be a threat to social agency, autonomy, and social trust.

IV.1. From Privacy to Autonomy and Social Agency - Shifting the Focus

IV.1.a. Traditional concerns regarding consent, privacy, and information sensitivity

When it comes to the specific case of the CA-Facebook scandal, first of all, there is an unquestionable breach of privacy as 87 million Facebook users' data was used without their knowledge and consent for the purpose of psychographically profiling and micro-targeting them with tailored messaging in favor of Trump's presidential campaign (Hu 2020). In fact, the developer of the personality-test app deployed on Facebook, Alexander Kogan, claimed that the data would be used for academic purposes only (ibid., 2). CA maintained that it had procured the data legally from Facebook and from commercial data brokers and played down the important role of the data in the company's project (ibid.). However, this was contradicted by whistleblower Wiley, who claimed that CA's algorithmic infrastructure was built on Facebook data (Cadwalladr & Graham-Harrison 2018). While the relevant regulatory agency of the US, the Federal Trade Commission (FTC), never went into litigation against Facebook, the unprecedented $5 billion penalty it imposed on the SMPF and the FTC's order regarding Facebook's handling of data and its transparency with

regard to it put aside doubts about the breach in privacy caused by the SMP. CA itself filed for bankruptcy in 2018 amid the scandal. As a result, the settlement from the FTC-CA lawsuit merely ordered Kogan and CEO Nix to refrain from making false statements about the use and procurement of the data and to delete all personal information obtained through the app and all "products" (including the algorithms) constructed on the basis of it (Federal Trade Commission 2019).

Thus, the fact that CA and Facebook illicitly procured data was the main issue of the settlement. What remained unaddressed, however, was the way in which personal- but also *non*-personal data (i.e., data through which the user could not be identified directly, e.g., her likes on Facebook) had been intricately processed through AI into psychological profiles on the basis of which users were targeted. As such, the FTC's settlements drew sharp criticism, including from within the Commission (Hu 2020, 3). This raises the question as to what is wrong about the micro-targeting methods CA employed? As Benn and Lazar (2022) write:

> Critics of Automated Influence argue that it relies on invasive inferences from data that is *illicitly acquired*, thereby delivering excessively targeted interventions that covertly shape people's beliefs, desires, and behaviour for exogenous ends. (Benn & Lazar 2022, 128)

Benn and Lazar point out that a prominent aspect of critiques of Automated Influence underline the way in which the data used for its processes was acquired. These critiques pertain to a rather developed body of scholarly work. Thus, one answer to the question regarding what is ethically wrong about psychological profiling reverts to an argument associated with the right to privacy. This argument emphasizes the sensitivity of the information that can be acquired by the technologically advanced methods of data analytics. Such sensitive information may include religious beliefs, political leanings, sexuality, and all other information a person may reasonably not want other people to know about her (Benn & Lazar 2022, 128). She would not want these things to be known about her as it could provide a means for potential wrong-doers to, for example, blackmail her, an employer to discriminate against her, or, as in the case of the CA-Facebook scandal, to be unknowingly targeted on the basis of psychological and correlational inferences without her consent. It could seem as though the former two hypotheticals are more unequivocally morally wrongful than the latter; blackmail involves a perpetrator well aware of what she is doing to the

victim as she infers how sensitive the information which she possesses about the victim is for the latter and uses this inference to extort something of value from her. Unfounded job discrimination is also clearly wrong, as the candidate finds herself unfairly prevented from getting an opportunity of potentially existential, basic need to her.

However, the critique of Automated Influence presented in this thesis goes beyond considering the way in which it achieves access to data and whether that data is personal or not and/or sensitive or not. In the CA case, if the inferences made about the millions of users were of sensitive nature, they were not made public, nor arguably used for such harmful purposes as extortion or discrimination. What would be wrong about CA's campaign if it had obtained its data in legal ways? By now it should be clear that this thesis' argument posits, first of all, that the wrong of Automated Influence goes beyond the issue of privacy. It even goes beyond its actual efficacy in modifying people's behavior based on constructing psychological profiles. Automated Influence inherently represents an attempt at undermining people's self-government by trying to interfere with their process of self-government, i.e., their autonomy. It relies not only on existing data formats (cf. Koopman 2022), it creates of the person - and imposes on it - a reconstructed identity. It imposes it on her through the digital, online, networked realm, notably SMP, where the algorithm determines what she is exposed to, at what moment, and for which ultimate purpose.

The means by which Automated Influence undermines people's autonomy are predictive calculations, i.e., algorithms. The latter can make inferences about what kind of messaging you would be susceptible to, including its framing, the topics, the tone, and where you're going to consume that, and how many times you need to be fed the effective messaging to change your perception/thinking on something (Cadwalladr & Graham-Harrison 2018). Based on calculative functions, the content in question is presented and potentially tailored to the SMP user's psychographically constructed profile in the hopes that it is efficacious in delivering the underlying purpose - to tune the user's motives and behavior according to the client's objective. Again, the effectiveness of the influencing campaign is questionable. But there is reason to believe that Automated Influence can be significant on a large scale (Benn & Lazar 2022). In the following section, I will consider the effects of Automated

Influence on an individual, rather than a group level. I will discuss how the data-driven algorithms trap people in a digital space that reinforces their confirmation bias and dis-incentivizes their re-evaluation of the reasons according to which they act. This leads us to considering, as Benn and Lazar emphasize, the impact of Automated Influence on the aggregate level.

IV.1.b. Intrapersonal autonomy - emotions and addiction in the 'attention economy'

While recommender systems and other types of Automated Influence target users with content based on what their behavioral data (expressed in clicks, views, and purchases, for example) reveals as *their* preferences, the variables and data used for the purposes of these algorithms have become more varied over time and, as the technology keeps developing, will continue to do so. For example, affective recommender systems are algorithms targeting SMP users with content that is expected to produce a certain emotional reaction (Mizgajski & Morzy 2018). As such, they can lean into the triggered emotional states that best serve their purpose. For recommender systems destined to maximize user engagement, it may be more beneficial to present the user with content generating a pleasant emotion. However, for other purposes, such as the political project underlying CA's campaign, it may be helpful to tap into emotions we may even intuitively deem as more problematic in terms of their instrumentalization. The content, or in the CA case specifically, the political ads deployed on the basis of psychographic profiles, for some people tapped into their fear. Platform users deemed "neurotic" (understood as a tendency to worry and anxiety), for example, were targeted by ads pointing to growing global instability and depicting chaotic scenes and juxtaposed the images and sentiments with a sense of stability and strength depicted at the view of the American flag and the campaigning politician (Merril & Goldhill 2020).

These examples show how Automated Influence not only uses inferences about preferences generated by behavioral data, but also seeks to tap into people's emotions as means to guide their pathway to action, as it is widely understood that they - along with preferences - play into behavior (Mizgajski & Morzy 2018). Emotions are something less controllable to the individual than her reasons. Certain, if not most emotions, cannot be reasoned with; they cannot be wished away or called upon according to the person's desire. In fact, they are partially responsible for steering her

desires. Thus, the express exploitation of human emotions for the purposes of influencing a person's behavior is problematic for her self-government.

Of course, the SMP user could theoretically only consume the kind of content she is looking for intentionally, but the interface of SMP is heavily structured by recommender systems that exist expressly to catch the attention of the user and keep her engaged with the algorithm-curated feed. Thus, the more powerful the algorithm (in terms of capturing the user's attention), the less likely she is to actively search for and navigate through content of her own intent choice. Because the algorithms structure content feeds in a way that can trigger addictive behavior (Bhargava & Velasquez 2021), the user can become 'trapped' in an informational ecology curated by opaque recommender systems. Thus, it becomes clear that the SMP user is incentivized not to take control over what she consumes, and even if she could be, the comfort of having entertaining content delivered automatically, or - her addiction to it, undermines the reality of autonomous choice.[2]

Besides the potential for addiction or simply the comfort within the automatically curated world of content, another layer to the issue is presented by the *type* of content that is deployed in the Automated Influence project. The nature of part of the content on SMP that is seemingly thriving in the information ecology created by recommender algorithms is misinformative, conspiracy adjacent, and politically divisive. I have mentioned that recommender algorithms may benefit from seeking to trigger a pleasant emotional response of the user to the content served to her. Such pleasure can, however, also be derived from confirming her bias. Moreover, tapping into her fear and fostering a discursive environment characterized by polarized tribalism can also prove satisfactory to the user - again - by confirming her bias. Since politics and the reality of the challenges we face locally and globally are complex in nature and people usually have a wish to understand them, simplified tales often have to rely on misinformation or conspiracy theories to make the message more digestible for the common person. These features make the content and message more salacious and are thus more effective in being promoted by algorithms and in convincing

[2] Of course, when consuming newspapers or television content, the consumer does not have immediate choice over what is presented to her either. That does not take away from the fact, however, that these media do not personalized the delivery of content and structure it according to what each individual consumer might prefer.

people, since the reality of many issues is far too complex to fully disentangle within a short video, written post, or political ad.

IV.1.c. Beliefs, reason(s), and interpersonal autonomy

Recommender systems not only prey on people's emotions. They also take into consideration everything else they are set out to infer about the SMP user in question. This can include inferred aspects of her personality, her sexuality, her religious and political beliefs, and so on. Predictive recommender algorithms tend to trap people in 'filter bubbles', reinforce their confirmation bias, and thus present a foreign influence in people's formation of their worldview. Even if the inferred personality traits, temperament, and political leaning of the user have been accurately captured by psychographics, the wrong about Automated Influence is that it then entraps her in her own worldview, preferences and dislikes, and political ideology. As such, Automated Influence represents a threat to personal development and incentivizes a lack of re-evaluation of one's beliefs. If it undermines the user's motive to re-evaluate her beliefs, it does the same to the reasons according to which she acts. As I have posited in the first chapter, the re-evaluation of reasons is a crucial element to autonomy and social agency. Agents, in order to be characterized as such, have to remain open to evaluating their reasons, especially within the social, deliberative space.

To be clear, any medium consumed by the person represents external ideas and interests that she exposes herself to: from newspapers to television programmes, any kind of expression of social, political, or economic views the 'consumer' opens herself up to has an impact on her worldview and, thus, her motives, reasons, and actions. However, these media do not continually and automatically reconfigure themselves on a personal basis. They are the same for everyone consuming them. Content (including political ads, videos, etc.) consumed through the SMP as a medium is not, however, presented to everyone in the same way. This is most obvious in the case of tailoring, but even the access to content (in terms of automatically being presented with it) determined by targeting renders each user's SMP interface unique and different from the next one's.

IV.2. From the Individual to the Group

In the beginning of this chapter, I conceded that the significance of Automated Influence on the individual level remains questionable. That is, while – if effective –

Automated Influence harms individuals' intrapersonal autonomy, the kind of harmful impact of concern to this thesis ought to be located on the collective, aggregate level, given that the primary ethical concern it deals with is of social, rather than individual nature. If we shift our inquiry to the aggregate level, meaning when we consider the sum of all potentially millions of people touched by the process, that the probability of its effectiveness, at least as a factor in a person's reasoning about a given topic (as for the CA case, a political election), plays a more significant role. The fact that an unprecedented number of people are targeted by Automated Influence campaigns inherently plays in its favor, as even a low success rate may generate impressive results.

There is good reason to believe that on an aggregate level, due to the sheer number of people being touched by Automated Influence processes, the latter's underlying objective can be effectively achieved (Benn & Lazar 2022). Benn and Lazar (2022) identify the phenomenon as "stochastic manipulation." Stochastic processes have "a random probability distribution or pattern that may be analysed statistically but may not be predicted precisely" (Oxford Languages 2023). As a consequence of the random probability pattern of the manipulation that Automated Influence exerts, they argue that its relevant impact is located at the aggregate, rather than the individual level. If the probability of effectively influencing an individual's behavior is rather low, then the effect of 'touching' a huge number of people (think of the CA case of 87 million SMP users) can still be significant.

Given the significance of Automated Influence on an aggregate, group level, and given the often divisive content riddled with disinformation or conspiracy theories (because these elements may be more exciting for users to consume), the impact on collective/group behavior and decisions is the greatest. This is of particular relevance for politics, for instance. Politics inherently concerns collective dynamics. In it, we mostly act through and affect change for groups brought and held together by shared interests. If Automated Influence reshuffles groups according to psychographics and exposes these newly formed groups to political discourse mediated by online content, what does this imply for political discourse and, ultimately, decision-making? As Benn and Lazar write:

> [S]tochastic manipulation preys on some pathologies of collective decision-making, in particular our failure to coordinate our actions with one another, and our propensity to realise tragedies of the commons. This is most obvious in the context of political decision-making—not just in elections, but more broadly when mobilising public support for or against particular policy proposals. In these contexts, the ability to sway a given group by a few percentage points, even a few fractions of a percentage point, can ultimately prove decisive (Heilman 2020). (Benn & Lazar 2022, 142)

Thus, even if individuals are rather unlikely to be swayed in direction of one course of action rather than another, due to the volume of people targeted be Automated Influence, it is very probable that some people are effectively influenced, which can have significant consequences for democratic deliberation and decision making (Benn & Lazar 2022). Because of the sensitivity of collective decisions and actions to the number of individuals in support or against them, Automated Influence could be decisive in steering democratic decision making.

IV.2.a. Social trust

In light of the stochastic manipulation perpetuated by Automated Influence, it can be reasonably argued that Automated Influence undermines social trust insofar as it provides legitimate grounds on the basis of which people mistrust the free, rational agency of one another. Thus, even if one were to question the entire idea of rational autonomy, the fact that it is what we base the potential for deliberation on (conceived as the social exchange, dialogue, and debate of reasons and objectives to attain collectively) makes the impact of Automated influence one that is harmful for the social relations of mutual recognition as free, rational, autonomous equals.

Even if a person remains steadfast in the idea that she herself is immune to manipulation, she cannot trust other people to be so too (ibid., 143). This represents a particular challenge to the social space for deliberation needed for social agency and autonomy. The lower person X's trust in the rational autonomy of her fellow members of society is, the lower the chance that democratic deliberation is effective. Even if the lack of autonomous rationality in light of Automated Influence is only perceived, the fact that it corrodes social trust implies that the exchange of reasons and deliberation are challenged, since both rely on this trust. In other words, Automated Influence undermines social trust, defined as the rational belief that fellow members of society are equally free, rational, autonomous agents.

This social trust is co-constitutive of social agency insofar as it grounds the social relations upon which the latter is conditioned. Social trust is a pre-condition for autonomous and rational free agency, because it sustains the social fabric underlying the relations upon which social agency relies.

IV.2.b. Social relations, normativity, and the deliberative-discursive space

There is a tangible risk that, through the personalization and individuation of the digital SMP space, the vital social-discursive space becomes fragmented. If we grant that SMP represent, to a degree, the new public square where people deliberate on socially and politically relevant issues, the fact that this space presents a different user experience to every individual affects the nature of the deliberation to be had there and possibly beyond the SMP space.

Automated Influence can overwhelm people with content they did not specifically set out to consume. The opacity of the algorithms it deploys to target people makes it impossible for them to know the reasons as to why they see certain content rather than other one. If, on an aggregate level, behavioral modification through affecting people's emotions, beliefs, and ultimately the reasons on which they act becomes (statistically) relevant, their process of developing reasons is interfered with. The fact that Automated Influence functions as a 'black box', i.e., a system whose internal workings remain opaque to those whom it affects and even – to a degree – to those who develop and deploy it (Hu 2020), renders people's reasoning more opaque too. If the content people consume shape their beliefs and worldviews and they do not even know on the basis of which inferences about them this content is shown to them, often even lacking the intentional awareness that it is presented to them for a data analytically calculated reason, this implies that people do not have enough insight into what is shaping their opinions and – ultimately – their motives for acting.

Let us recall at this point that the essential element for autonomy as a condition for social-agency claims lies in the social recognition of the reasons according to which an agent acts. By constantly being shown content leaning, for example, in the same political-ideological direction, the SMP user may find her beliefs and worldview altered (or hardened) and, as a result, so too the reasons upon which she acts. The calculative predictions behind the Automated Influence systems influence her without

her awareness, or, at the very least, without her knowing how they work and whether the reasons for which she might have a 'change of heart' derive from her own re-evaluations or were brought onto her by means of precise algorithmic targeting. As a result, the social space serving to ground reasons in normativity and to allow for deliberation becomes distorted by an external, powerful influence using AI to compute and predict human behavior, and which, if the technologies are to improve in coming years and the data flow to increase, will become better at approximating action to prediction. In a world of rapid technological progress in the domain of AI and given the digital transition that is now widely accepted as a political project emanating from the West to the entire world, it is more probable that Automated Influence will increase in effectiveness.

Thus, for the social-discursive space tasked with grounding the normativity of reasons, the problem posed by Automated Influence lies principally in the opacity of the process itself (from data collection to the algorithms used to target people), as well as - potentially - the objective behind it, in the case of third-party driven Automated Influence on SMP. More generally, the problem presents itself as a lack of awareness by the people who are subjected to Automated Influence. Not only are they unknowledgeable about how it works and for what purposes, they also - as a result - know less and less about the reasons behind them consuming certain content over other, because of the covert/implicit manner in which algorithms operate.

The nature of the contemporary attention economy produces incentives and Automated Influence techniques allow to get as deeply as possible into the minds of people in order to keep them engaged and nudge them toward profitable ends. In the case of recommender algorithms, Automated Influence seeks to boost user engagement on SMP. In the case of a third-party contractor like CA, who collaborated with an SMP, the purposes can be much bigger in conception, even if the technological means to achieve it remain the same. In addition, from the political to the commercial realm, as long as Automated Influence remains a service to be purchased, the projects and intentions guiding it could vary from harmless (in terms of the fundamental interests I am concerned with) to actually harmful. On the individual level, the wrongfulness of how Automated Influence functions lies in the fact that it preys on people's emotions and can even trigger addictive behavior. On the group

level, Automated Influence corrodes social trust necessary to sustain the kind of social relations presupposed for social agency, as well as the normative discursive space indispensable for autonomy. Moreover, the opacity of the processes and the general lack of awareness among people about Automated Influence is problematic for their autonomy, as they are kept in the dark about what it is about them that makes them be presented with said content. It is these inferences which are made about them by machine intelligence processes and which they have no access to our power in shaping that ultimately present an affront to their autonomy and social agency.

V. Addressing the Problem: From Top-Down to Horizontal Approaches

As noted in the beginning, this thesis sets out to formulate a political response to the impediments posed by Automated Influence to social agency. Before laying out the proper argument, I will give a brief overview of the existing political and legislative approaches to the issues posed by Automated Influence. Much of the regulation (implicitly) targeting Automated Influence and related processes (i.e., obtention, storage, and use of data; automated decision making through algorithms) specifically targets SMP/SMPF, since they represent the main host for Automated Influence systems. While forgoing any discussion as to the reasons behind it, Cioffi, Kenney, and Zysman (2022, 826) note that around 2019 to 2020, a considerable shift of approach toward regulating SMP took place. In my view, it is undeniable that the highly publicized CA scandal and especially Facebook's involvement in it contributed to nudge regulators in this direction. After the Facebook-Cambridge-Analytica scandal came to light, politicians and the public writ large were swift to condemn the rather straightforward breach in privacy rights and question the ethics of large-scale political influence campaigns in the context of novel technological means. The outcry was likely in part amplified by the already existing political tensions and polarization around Brexit and Trump's presidential campaign, both of which were objects of CA's Automated Influence campaigns.

As it became clear that SMPF do not only represent immensely powerful market actors, but also impact societies beyond economically relevant relevant factors, regulators started to shift their focus from exclusively targeting competitive concerns

toward a more holistic approach in confronting the disruptive impacts of SMP. In contrast to previous *ex post* competition law, regulators have moved towards *ex ante*, proscriptive and/or prescriptive rules of general application with regards to a variety of categories of SMPF behavior "to prevent categorical forms of harm" (ibid., 827). The European Union (EU) is at the forefront of regulating the digital online sphere, which is why even though this thesis leans on the CA case with regard to the Trump presidential campaign, I will examine the EU's approach rather than the US'.

There is a vast variety of purposes served by Automated Influence, a significant part of which could be reasonably assessed as harmless, if not beneficial in terms of rendering the SMP service more efficient and user-friendly. As a result, targeting Automated Influence as such becomes a complex task, as it varies in functions and underlying objectives.

V.1. Existing legislative responses

For a long time, existing legislation relevant to Automated Influence primarily focused on privacy concerns and thus targeted the processes related to the obtention, storage, and use of data, especially of sensitive nature (such as personal or financial data). The GDPR also largely focuses on privacy, but it explicitly set out to target data processes more holistically, notably with regard to the "fundamental rights and freedoms of natural persons". However, as will be clear by my critique of the GDPR, the regulation remains trapped in the normative logic of privacy and subsequently delivers an excessively methodologically individualist response to the ethical challenges of Big Data and Automated Influence, as a subset of practices under the umbrella of the latter.

V.1.a. The EU's General Data Protection Regulation

The regulatory move beyond concerns of privacy was instantiated by the EU's General Data Protection Regulation (GDPR), which came into effect on May 25, 2018, approximately two months after the revelations of the CA-Facebook scandal and only a few days after CA had declared bankruptcy. While the GDPR explicitly aims at preserving personal privacy, it also sets out to target mass surveillance (Andrew & Baker 2021, 570). Moreover, the GDPR seeks to address the rising power of Big Data players, including SMPF, the market in which they compete, and concerns for the consumer (ibid.). It has introduced "changes related to remedies,

fees, restrictions, enforcement, and organizational practices for establishing consent and trust around the collection and general use personal data" (ibid., 568).

Addressing data lifecycles and value chains

Thus, the GDPR addresses processes related to data obtention, storage, use, and transparency. First of all, through the principle of "data minimization", the GDPR stipulates the requirement of limiting data collection to what is necessary for the purpose defined by the data collector. With regard to the latter, the GDPR (Article 5.1b) legally requires the purpose(s) of the data collection and processing to be "specific, explicit, and legitimate" (ibid., 571). As a result, organizations (both in the private and public sector) active in data collection, processing, and exchange are legally obligated to be transparent toward the data subject/person when it comes to how her data will be used. In addition, the GDPR imposes a time limit to the storage of data after which the organization is required to undergo a periodic review concerning the data or to erase it. If the data already obtained and stored is to be reused for another purpose, the organization must notify and inform the data subjects in question about this (ibid.). The legitimacy of the purpose(s) behind data processing is presumably conceived in terms of its legality.

In addition to stipulating conditions to data processing, the GDPR explicitly distinguishes between different categories of data. Personal data denotes any data which "relates to, or accurately describes, some aspect of a living, identifiable individual" (ibid.), ranging from a person's name and location to her IP address, for example. Within the category of personal data, the GDPR pays special attention to sensitive data, which, as already pointed out in the previous chapter, reveals aspects of the person that are especially sensitive in terms of the potential for harmful use thereof. As such, sensitive data includes information such as the ethnic or racial origins of a person, her political opinions, her beliefs of philosophical or religious nature, trade-union membership, and information about her health and sexual preferences (ibid.). Processing of sensitive data is only allowed on the condition of obtaining the data subject's explicit consent, for purposes deemed of "substantial public interest or public safety", or "special medical treatment" (ibid.). This shows how the GDPR focuses on unequivocal, intentional forms of harm perpetrated by

other individuals or organizations, especially with regards to the right to privacy, rather than having broader, social impacts in mind.

A lagging response to automated decision-making

Of more immediate relevance to Automated Influence processes is the legal framework the GDPR provides for the conditions under which automated decision-making is allowed. Here, again, it stipulates as a condition the data subject's explicit consent to being subjected to these techniques. Moreover, it obligates organizations to provide the data subject with "meaningful information" about the logic underlying the algorithm fulfilling the task and the potential consequences of the data processing for her, if she requests this information (ibid.). This may be an adequate response when it comes to automated decision-making within domains that - from the data subject's perspective, are unequivocally impactful on her livelihood. An example would be when a person takes out a loan from an online bank, which calculates the parameters of the financial instrument, such as the interest rate (or whether she has access to a loan at all). However, based on the general lack of awareness about automated decision-making as a whole, it is highly questionable that most, if even a significant chunk of data subjects would request such information, especially when it comes to more seemingly innocuous types of automated decisions, including the ones involved in Automated Influence systems (e.g., recommender systems).

The GDPR also sets requires human intervention where decisions produce legal affects or significantly affect the individual "in a similar way" (European Commission 2023a). Legal effects are impacts on the legal rights of persons. Other significant effects are described as influences on a person's "circumstances, behavior, or choices" (ibid.). On the 'Q&A' website regarding the question "Can I be subject to automated individual decision-making, including profiling?", the European Commission states the aforementioned example of online credit applications as bearing significant impact (ibid.). It is rather clear that significant impacts are conceived as influencing the legal-political and material conditions of the person. While Article 22.3 specifies that "the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests", it becomes clear that this protection of rights, freedoms, and (especially) legitimate interests are conceived rather narrowly, since the text immediately follows up by

stipulating as a minimal necessary condition "the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision" (European Parliament & Council of the European Union 2016).

As such, the GDPR formulates the problem purely in terms of the automated nature of the decision making. While it is true that the automated nature of the decision-making (or, according to my focus, influence) is problematic, especially with regards to its opacity, the effect produced by it is not guaranteed to be nullified as a result of the required human sanctioning of the process. What the GDPR sets out to do through the requirement of human intervention is to prevent egregious cases of discrimination and other effects impeding on people's "legal rights." In those cases, at least if there is an entire legal motion to contest the decision, requiring human intervention as the ultimate sanctioner can reasonably be expected to work.

Incentivization for of data processing as a result of hyperfocusing on privacy

In addition, by doing away with the purpose specification requirement for pseudonymized data, the GDPR has institutionalized a hugely significant caveat with regards to data processing and the previously stipulated conditions surrounding it. Pseudonymized data is data which - by itself - does not render the data subject identifiable. Only in combination with additional data does pseudonymized data allow to identify the data subject (ibid., 572-73). This caveat allowing for pseudonymized data to be used beyond the purposes it was initially collected for is intended to serve as an incentive for organizations to use pseudonymized data so as to preserve the privacy of data subjects and to avoid targeting individuals by singling them out (ibid., 573). As such, Facebook providing CA with millions of users' data is considered illegal under the GDPR, since the data used was not pseudonymized (nor anonymized) and explicitly targeted SMP users on the basis of their individually inferrable characteristics. That said, this exchange of data would not have been legal if the latter had been pseudonymized, since neither Facebook nor CA specified the purpose for which it was used and, thus, did not ask for the data subjects' consent.

However, as Andrew and Baker argue, the fact that pseudonymized data is allowed to be used for purposes beyond those it was initially collected for (including selling it to third parties) actually promotes the commodification and exchange of de-identified data, as well as of the predictive models used to interpret it. As Shoshana Zuboff's

theory of surveillance capitalism explains, Big Data industries are not primarily interested in targeting and observing the identifiable individual. Rather, non-identified data serves as "raw material" that is processed through AI technologies and can generate immense economic value because of its predictive capacities (ibid., 574).[3] The GDPR would be more effective in curtailing mass surveillance as a whole if it remained firm on the principles of purpose specification and data minimization (ibid., 573).

The inadequacy of individual consent for aggregate impact

Thus, while the GDPR addresses individuals' privacy concerns, it does not regulate mass surveillance interested in making inferences about groups, nor the purposes for which these inferences may be used. This observation serves to re-emphasize the significance of Big Data and Automated Influence for group dynamics and behavior and restates the importance of focusing relevant inquiries on the aggregate level.

According to Andrew and Baker (2021, 570) "the GDPR represents a significant intensification of the legal environment in that it is binding for all member states, and includes the rights to erasure and access; requirements for affirmative consent; and the ability to levy heavy fines." Thus, while it has inaugurated an institutional framework for the regulation of data usage, storage, and flows, it still operates on the presumption that individual consent is a legitimate sufficient condition for many of the same data processing ventures to continue to operate in the digital online realm.

The problem with individual consent when it comes to Automated Influence practices is that the public is far from aware about the potential implications of Automated Influence. As studies have shown (cf. Hinds, Williams & Joinson 2020), even when they are, they are reticent in opting out of the technologies in question. If they are aware and do not deem themselves but others susceptible to online manipulation, individual consent becomes irrelevant, as people can only choose for themselves whether they opt in or out of Automated Influence regimes and not for others. The issue posed by the corrosion of social trust emanating from Automated Influence points to the problem of methodological individualism when it comes to ethically mediating the impact Automated Influence. Individual consent cannot be the answer

[3] Thus, the GDPR can also be said to fail with regards to its objective to curtail the market power of influential Big Data industries (Andrew & Baker 2021, 574).

to the question as to if and how we deem the Automated Influence as legitimate on the group level, where legitimacy is conceived more broadly than (implicitly) in the GDPR.

V.1.b. The Digital Services Act

In addition to the GDPR, another piece of EU legislation has attracted substantial attention as "the furthest reaching expansion of platform regulation in the OECD nations to date" (Cioffi et al. 2022, 828). If the GDPR targets data across the board, the Digital Services Act (DSA) targets platforms specifically with the objectives to protect the consumer and her personal privacy, upping platform transparency with regards to their terms and conditions, market practices, and their handling of personal and financial data (ibid., 829).

In addition, the DSA bans online targeted advertising based on profiling children, as well as based on sensitive data, "such as ethnicity, political views or sexual orientation" (European Commission 2023b). This is, without a doubt, a critical move forward in the regulation of online platforms, including SMP, as it outlaws (in principle) forms of advertising such as the one promoted by CA on Facebook. It categorically condemns any advertising, however innocuous its object may be, if it is based on targeting methods processing data containing sensitive information. Given that the DSA has yet to enter into force, it remains to be determined whether such stipulations do not fall short (as some of the GDPR do) with regards to the technological capacities of platform firms and Big Data industries in general. In other words, there is a possibility that data analytics technologies will enable ways to work around the sensitive-data label and thus evade this particular DSA provision. Nonetheless, as of now, it does not seem as though the DSA would introduce a clause similar to the pseudonymization caveat of the GDPR.

Moreover, the DSA also requires platform companies to be transparent with regards to their policies on algorithmic decision-making on their platform, but the text remains vague with regard to the adequate level of transparency (European Commission 2023b). Since platforms and algorithms vary immensely in form, the omission of clear minimum criteria for transparency produces uncertain legal effects. Transparency requirements are also stipulated for online advertising. Any user must

be able to identify a given ad and access information on the parameters of the targeting method used for the ad.

While these are significant steps towards allowing platform users to become more aware about automated decision-making as it relates to the platform interface, such provisions fall short of tackling the implications of large-scale Automated Influence. Again, the onus is put onto the individual to inform and educate herself, which can be beneficial to her own awareness about methods of targeting, but does not preclude stochastic manipulation. It is rather unlikely that most, if a significant chunk of people will consistently consult information about why they are shown certain ads. Even if they do, the information revealed about the targeting process is rather sparse and will likely fail to convey an adequate vision of the enormous scale of data processing and the aggregate impact of algorithms for Automated Influence.

Both under the GDPR and the DSA, more straightforwardly unethical Automated Influence projects are rendered illegal. This is a good start, but fails to address concerns related to the impact of the more accepted Automated Influence systems such as recommender algorithms governing SMP interfaces and targeting people with content. What the ex ante/general applicability approach of administrative regulation gets right is that it posits normative objectives as legally prior to all kinds of business behavior. However, when it comes to the issue of Automated Influence and its wide variety of forms and objectives, regulation is unlikely to adequately target impediments to social agency and trust propagated by more seemingly innocuous forms of Automated Influence, like recommender systems.

V.2. Towards a Horizontal Approach

While there certainly is merit to the fact that political figures are taking action in the light of new technological processes considered by many to interfere with social and political discourse and, ultimately, decision-making, there are, as I have pointed out, many shortcomings to the regulatory approach to tackling the wrong of Automated Influence I have laid out in the previous chapter.

Notwithstanding the more typically evoked problems with regards to regulation, such as rent-seeking behavior by the state, excessive bureaucratization, slow processing

(Hazlett 2022), for the regulation of Automated Influence and SMPF, the hindrance to effective regulation lies in the nature of the issue at hand. As I have pointed out, a major aspect involved in the harm perpetrated by Automated Influence is its black-box operations, meaning its opaque system and what the latter implies for the general public's awareness about who could seek to influence their behavior, for what ends, based on which inferences about them as individuals, etc.

In response to the issue of opacity, which, in terms of state regulation, may only be fully addressed by the state's acquisition of SMPF, one could argue in favor of institutionalizing a more robust set of ethical principles to be followed by private companies themselves. Unless SMPF becomes state-owned enterprises (which would raise a whole set of ethical and political issues by itself, such as the risk of authoritarian regimes controlling an enormous chunk of the information economy), the onus would be on the private companies themselves in conducting their operations in ways that do not conflict with ethical principles, such as the respect of people's fundamental interests. Such an approach has been laid out by Abraham Singer, who has argued in favor in recalibrating the "moral division of labor" so as to oblige businesses to comply with justice-based demands when such harms are perpetrated by only following the principle of efficiency (Blunden 2022). However, such approaches to tackling the problem at hand are rather unpromising in practical terms and raise a whole set of problems: What kind of ethical principles ought to guide private businesses and who are they determined by? How is this compliance ultimately enforced? In the light of the question of enforcement and given that the state retains the monopoly of violence, would such a business-ethics approach not ultimately revert to state administrative regulation?

Ultimately, while it would in many cases be morally desirable to have private companies be guided by more than just concerns for efficiency, the practical execution of such a business ethics is improbable, if not impossible under a capitalist economy. While business-ethics approaches seek to circumvent problems associated with state regulation, they are difficult to implement in practice. What, then, could pose a more suitable solution for the issue at hand?

Given that a major aspect of the problem at hand derives from the opacity of Automated Influence and a lack of awareness of the general public regarding its existence and operations, I argue that the most reasonable approach to tackling the issue is to educate people about it. This task should fall on society as a whole, rather than just the state, although the latter ought to be involved insofar as public educational curricula are concerned. Emerging, complex technologies are transforming people's lives. Besides the potentially socially corrosive impacts of SMP I have focused on, widely available AI tools are transforming the way children are educated and challenging academic and artistic authenticity (think of the current controversy regarding the AI language model ChatGPT), and automation is squeezing the labor market. These are only a few examples of the disruptive, transformative impact of contemporary technologies but they serve to point to the fact that human societies are facing unprecedented opportunities, but also challenges in the face of technological progress. While the consequences of automation for the availability of 'unskilled' (and potentially even 'skilled') jobs is a straightforwardly political challenge, the other examples point to a possibility of mitigating potential damages by educating people about the technologies in question and the risks, as well as the benefits they are associated with.

When it comes to Automated Influence specifically, people should be educated about the kind of media through which it operates (mostly online, digital interfaces), in which specific operations (e.g., content recommendation, advertisements), how, specifically on SMP, it might 'trap' them in filter bubbles affecting their desires, beliefs, and worldviews. This cannot guarantee that its harmful impacts are completely done away with, but it seems to be the most sustainable and adequate strategy to mitigate them. The impact of Automated Influence on the social fabric can, in the long run, only be addressed by raising awareness about the issue so as to give people the opportunity to make more genuinely informed choices about their participation in the attention economy. Moreover, if people are more educated about how online discourse mediated to them through Automated Influence might affect their beliefs and motives, they might reflect more intensely on their worldviews, be more attentive to the kind of social and political messaging they are exposed to, and gain more insight into their reflective processes as a whole. Only through education can intra- and interpersonal autonomy flourish and be protected from emerging modes

of potent and technologically complex modes of influence. Hence, it also serves to restore social agency and social trust, as people can count on their fellow members of society having an enhanced understanding of what may influence, if not manipulate them.

VI. Conclusion

In his interview with *The Guardian*, Cambridge Analytica whistleblower Christopher Wiley formulates the problem posed by Automated Influence to society in a way that aptly reflects the central critique of this thesis. As he states:

> "Instead of standing in the public square and saying what you think and then letting people come and listen to you and have that shared experience as to what your narrative is, you are whispering into the ear of each and every voter and you may be whispering one thing to this voter and another thing to another voter. [...] We risk fragmenting society in a way where we don't have any more shared experiences and we don't have any more shared understanding. If we don't have any more shared understanding, how can we be a functioning society? (Christopher Wiley in Cadwalladr & Graham-Harrison 2018)

Together, a shared space of common understandings and social trust provide a necessary basis for the respect of people's social agency. However, it is not only the rather straightforwardly wrongful uses of Automated Influence exemplified by the Facebook-CA scandal that should attract our attention. Rather, seemingly innocuous recommender systems, the algorithms governing content suggestions and structuration on SMP, also impact people's beliefs, worldviews, reasoning, and action. While the efficacy of Automated Influence is a matter for empirical research, the fact that members of society now have to operate in a social space where the trust in their fellow citizens' autonomy and rationality is more under attack than ever already has important implications for the social space of deliberation and discourse on which the exercise of social agency relies.

While state regulation targeting processes and practices related to Automated Influence have been successful in sheltering individual privacy and increasingly set out to defend people against straightforward wrongdoings, it is less clear it will ever grasp the true challenge posed by Automated Influence. Countering it ought to be an

equally bottom-up as a top-down exercise. In the long run, it is only by becoming educated about the phenomenon that people can reclaim their agency and autonomy.

Will the social fabric holding societies together be torn apart by the increasing fragmentation brought about by Automated Influence and the withering away of social trust, which renders democratic politics and consensus-based decision-making increasingly difficult? While only time will tell how societies will transform under Automated Influence, in light of this thesis' critique, it seems clear that it represents yet another factor of uncertainty for society and politics, in addition to the economic, environmental, and geopolitical pressures and shifts emblematic of the first quarter of the twenty-first century.

Bibliography

Anderson, Joel. 2003. "Autonomy and the Authority of Personal Commitments: From Internal Coherence to Social Normativity." *Philosophical Explorations* 6 (2): 90–108. https://doi.org/10.1080/10002003058538742.

Andrew, Jane and Max Baker. 2021. "The General Data Protection Regulation in the Age of Surveillance Capitalism." *Journal of Business Ethics* 168 (3): 565-578. https://doi.org/10.1007/s10551-019-04239-z.

Benn, Claire, and Seth Lazar. 2022. "What's Wrong with Automated Influence." *Canadian Journal of Philosophy* 52 (1): 125–48. https://doi.org/10.1017/can.2021.23.

Bhargava, Vikram R., and Manuel Velasquez. "Ethics of the Attention Economy: The Problem of Social Media Addiction." *Business Ethics Quarterly* 31 (3): 321–59. https://doi.org/10.1017/beq.2020.32.

Blunden, Charlie. 2022. "Between Market Failures and Justice Failures: Trade-Offs Between Efficiency and Equality in Business Ethics." *Journal of Business Ethics* 178 (3): 647-660. https://doi.org/10.1007/s10551-021-04767-7.

Buss, Sarah and Andrea Westlund. 2018. "Personal Autonomy." In *Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. https://plato.stanford.edu/archives/spr2018/entries/personal-autonomy/.

Cadwalladr, Carole and Emma Graham-Harrison. "Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach." *The Guardian*, March 17, 2018. https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election.

Cioffi, John W., Martin F. Kenney, and John Zysman. 2022. "Platform Power and Regulatory Politics: Polanyi for the Twenty-first Century." *New Political Economy* 27 (5): 820-836. https://doi.org/10.1080/13563467.2022.2027355.

European Commission. 2023. "Can I be subject to automated individual decision-making, including profiling?" Accessed July 25, 2023. https://commission.europa.eu/law/law-topic/data-protection/reform/rights-citizens/my-rights/can-i-be-subject-automated-individual-decision-making-including-profiling_en.

European Commission. 2023. "Questions and Answers: Digital Services Act." Last modified April 25, 2023. https://ec.europa.eu/commission/presscorner/detail/en/QANDA_20_2348.

European Parliament and Council of the European Union. 2016/679. *Regulation of the European Parliament and of the Council of 27 April 2016 on the protection of natural*

persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).

European Parliament. 2018. *European Parliament resolution of 25 October 2018 on the use of Facebook users' data by Cambridge Analytica and the impact on data protection.* (2018/2855(RSP)).

Federal Trade Commission. 2019. "FTC Grants Final Approval to Settlement with Former Cambridge Analytica CEO, App Developer over Allegations they Deceived Consumers over Collection of Facebook Data." Published December 18, 2019. https://www.ftc.gov/news-events/news/press-releases/2019/12/ftc-grants-final-approval-settlement-former-cambridge-analytica-ceo-app-developer-over-allegations.

Hern, Alex. "Cambridge Analytica: how did it turn clicks into votes?" *The Guardian*, May 6, 2018. https://www.theguardian.com/news/2018/may/06/cambridge-analytica-how-turn-clicks-into-votes-christopher-wylie.

Hill, Seth J., Daniel J. Hopkins, and Gregory A Huber. 2021. "Not by Turnout Alone: Measuring the Sources of Electoral Change, 2012 to 2016." *Science Advances* 7 (17). https://doi.org/10.1126/sciadv.abe3272.

Hinds, Joanne, Emma J. Williams, and Adam N. Joinson. 2020. "'It wouldn't happen to me': Privacy concerns and perspectives following the Cambridge Analytica scandal." *International Journal of Human-Computer Studies* 143 (1): 1-14. https://doi.org/10.1016/j.ijhcs.2020.102498.

Hu, Margaret. 2020. "Cambridge Analytica's Black Box." *Big Data & Society* 7 (2): 1-6. https://doi.org/10.1177/2053951720938091.

Koopman, Colin. 2022. "The Political Theory of Data: Institutions, Algorithms, & Formats in Racial Redlining." *Political Theory* 50 (2): 337-361. https://doi.org/10.1177/00905917211027835.

Lewis, Paul and Paul Hilder. "Leaked: Cambridge Analytica's blueprint for Trump victory." *The Guardian*, March 23, 2018. https://www.theguardian.com/uk-news/2018/mar/23/leaked-cambridge-analyticas-blueprint-for-trump-victory.

Merill, Jeremy B. and Olivia Goldhill. 2020. "These are the political ads Cambridge Analytica designed for you." *Quartz*. Published January 10, 2020. https://qz.com/1782348/cambridge-analytica-used-these-5-political-ads-to-target-voters.

Noor, Poppy. "There are plenty more like Cambridge Analytica. I know – I've used the data." *The Guardian*, March 23, 2018. https://www.theguardian.com/commentisfree/2018/mar/23/plenty-more-like-cambridge-analytica-data-facebook.

Osborne, Hillary. “What is Cambridge Analytica? The firm at the centre of Facebook's data breach.” *The Guardian*, March 18, 2018. https://www.theguardian.com/news/2018/mar/18/what-is-cambridge-analytica-firm-at-centre-of-facebook-data-breach.

Schuppert, Fabian. 2013. “Distinguishing Basic Needs and Fundamental Interests.” *Critical Review of International Social and Political Philosophy* 16 (1): 24–44. https://doi.org/10.1080/13698230.2011.583532.

Zuboff, Shoshana. 2019. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. New York: Public Affairs.

Zuboff, Shoshana. 2015. “Big Other: Surveillance Capitalism and the Prospects of an Information Civilization.” *Journal of Information Technology* 30 (1): 75-89. https://doi:10.1057/jit.2015.5.