# Chapter 5
# Manipulation


**Abstract** The concern that artificial intelligence (AI) can be used to manipulate individuals, with undesirable consequences for the manipulated individual as well as society as a whole, plays a key role in the debate on the ethics of AI. This chapter uses the case of the political manipulation of voters and that of the manipulation of vulnerable consumers as studies to explore how AI can contribute to and facilitate manipulation and how such manipulation can be evaluated from an ethical perspective. The chapter presents some proposed ways of dealing with the ethics of manipulation with reference to data protection, privacy and transparency in the of use of data. Manipulation is thus an ethical issue of AI that is closely related to other issues discussed in this book.

**Keywords** Right to life · Safety · Security · Self-driving cars · Smart homes · Adversarial attacks · Responsibility · Liability · Quality management · Adversarial robustness

## 5.1 Introduction

In the wake of the 2016 US presidential election and the 2016 Brexit referendum it became clear that AI had been used to target undecided voters and persuade them to vote in a particular direction. Both polls were close, and a change of mind by a single-digit percentage of the voter population would have been enough to change the outcome. It is therefore reasonable to state that these interventions led by artificial intelligence (AI) played a causal role in the ascent of Donald Trump to the American presidency and the success of the Brexit campaign.

These examples of the potential manipulation of elections are probably the most high-profile cases of human action being influenced using AI. They are not the only ones, however, and they point to the possibility of much further-reaching manipulation activities that may be happening already, but are currently undetected.



B. C. Stahl et al., *Ethics of Artificial Intelligence*,
SpringerBriefs in Research and Innovation Governance,
https://doi.org/10.1007/978-3-031-17040-9_5

## 5.2 Cases of AI-Enabled Manipulation

### *5.2.1 Case 1: Election Manipulation*

The 2008 US presidential election has been described as the first that "relied on large-scale analysis of social media data, which was used to improve fundraising efforts and to coordinate volunteers" (Polonski 2017). The increasing availability of large data sets and AI-enabled algorithms led to the recognition of new possibilities of technology use in elections. In the early 2010s, Cambridge Analytica, a voter-profiling company, wanted to become active in the 2014 US midterm election (Rosenberg et al. 2018). The company attracted a $15 million investment from Robert Mercer, a Republican donor, and engaged Stephen Bannon, who later played a key role in President Trump's 2016 campaign and was an important early member of the Trump cabinet. Cambridge Analytica lacked the data required for voter profiling, so it solved this problem with Facebook data (Cadwalladr and Graham-Harrison 2018). Using a permission to harvest data for academic research purposes that Facebook had granted to Aleksandr Kogan, a researcher with links to Cambridge University, the company harvested not just the data of people who had been paid to take a personality quiz, but also that of their friends. This allowed Cambridge Analytica to harvest in total 50 million Facebook profiles, which allowed the delivery of personalised messages to the profile holders and also – importantly – a wider analysis of voter behaviour.

The Cambridge Analytica case led to a broader discussion of the permissible and appropriate uses of technology in Western democracies. Analysing large datasets with a view to classifying demographics into small subsets and tailoring individual messages designed to curry favour with the individuals requires data analytics techniques that are part of the family of technologies typically called AI.

We will return to the question of the ethical evaluation of manipulation below. The questions that are raised by manipulation will become clearer when we look at a second example, this one in the commercial sphere.

### *5.2.2 Case 2: Pushing Sales During "Prime Vulnerability Moments"*

Human beings do not feel and behave the same way all of the time; they have ups and downs, times when they feel more resilient and times when they feel less so. A 2013 marketing study suggests that one can identify typical times when people feel more vulnerable than usual. US women across different demographic categories, for example, have been found to feel least attractive on Mondays, and therefore possibly more open to buying beauty products (PHD Media 2013). This study goes on to suggest that such insights can be used to develop bespoke marketing strategies. While the original study couches this approach in positive terms such as "encourage"

and "empower", independent observers have suggested that it may be the "grossest advertising strategy of all time" (Rosen 2013).

Large internet companies such as Google and Amazon use data they collect about potential customers to promote goods and services that their algorithms suggest searchers are in need of or looking for. This approach could easily be combined with the concept of "prime vulnerability moments", where real-time data analysis is used to identify such moments in much more detail than the initial study.

The potential manipulation described in this second case study is already so widespread that it may not be noticeable any more. Most internet users are used to being targeted in advertising.

The angle of the case that is interesting here is the use of the "prime vulnerability moment", which is not yet a concept widely referred to in AI-driven personal marketing. The absence of a word for this concept does not mean, however, that the underlying approach is not used. As indicated, the company undertaking the original study couched the approach in positive and supportive terms. The outcome of such a marketing strategy may in fact be positive for the target audience. If a person has a vulnerable moment due to fatigue, suggestions of relevant health and wellbeing products might help combat that state. This leads us to the question we will now discuss: whether and in what circumstances manipulation arises, and how it can be evaluated from an ethical position.

## 5.3 The Ethics of Manipulation

An ethical analysis of the concept of manipulation should start with an acknowledgement that the term carries moral connotations. The Cambridge online dictionary offers the following definition: "controlling someone or something to your own advantage, often unfairly or dishonestly" (Cambridge Dictionary n.d.) and adds that it is used mainly in a disapproving way. The definition thus offers several pointers to why manipulation is seen as ethically problematic. The act of controlling others may be regarded as concerning, especially the fact that it is done for someone's advantage, which is exacerbated if it is done unfairly or dishonestly. In traditional philosophical terms, it is Kant's prominent categorical imperative that prohibits such manipulation on ethical grounds, because one person is being used solely as a means to another person's ends (Kant 1998: 37 [4:428]).

One aspect of the discussion that is pertinent to the first case study is that the manipulation of the electorate through AI can damage democracy.

> AI can have (and likely already has) an adverse impact on democracy, in particular where it comes to: (i) social and political discourse, access to information and voter influence, (ii) inequality and segregation and (iii) systemic failure or disruption. (Muller 2020: 12)

Manipulation of voters using AI techniques can fall under heading (i) as voter influence. However, it is not clear under which circumstances such influence on voters would be illegitimate. After all, election campaigns explicitly aim to influence voters and doing so is the daily work of politicians. The issue seems to be not so much the fact that voters are influenced, but that this happens without their knowledge and maybe in ways that sidestep their ability to critically reflect on election messages. An added concern is the fact that AI is mostly held and made use of by large companies, and that these are already perceived to have an outsized influence on policy decisions, which can be further extended through their ability to influence voters. This contributes to the "concentration of technological, economic and political power among a few mega corporations [that] could allow them undue influence over governments" (European Parliament 2020: 16).

Another answer to the question why AI-enabled manipulation is ethically problematic is that it is based on privacy infringements and constitutes surveillance. This is certainly a key aspect of the Cambridge Analytica case, where the data of Facebook users was harvested in many cases without their consent or awareness. This interpretation would render the manipulation problem a subproblem of the broader discussion of privacy, data protection and surveillance as discussed in Chap. 3.

However, the issue of manipulation, while potentially linked with privacy and other concerns, seems to point to a different fundamental ethical concern. In being manipulated, the objects of manipulation, whether citizens and voters or consumers, seem to be deprived of their basic freedom to make informed decisions.

Freedom is a well-established ethical value that finds its expressions in many aspects of liberal democracy and forms a basis of human rights. It is also a very complex concept that has been discussed intensively by moral philosophers and others over millennia (Mill 1859; Berlin 2002). While it may sound intuitively plausible to say that manipulating individuals using AI-based tools reduces their freedom to act as they normally would, it is more difficult to determine whether or how this is the case. There are numerous interventions which claim that AI can influence human behaviour (Whittle 2021), for example by understanding cognitive biases and using them to further one's own ends (Maynard 2019). In particular the collecting of data from social media seems to provide a plausible basis for this claim, where manipulation (Mind Matters 2018) is used to increase corporate profits (Yearsley 2017). However, any such interventions look different from other threats to our freedom to act or to decide, such as incarceration and brainwashing.

Facebook users in the Cambridge Analytica case were not forced to vote in a particular way but received input that influenced their voting behaviour. Of course, this is the intended outcome of election campaigns. Clearly the argument cannot be that one should never attempt to influence other people's behaviour. This is what the law and, to some extent, ethics do as a matter of course. Governments, companies and also special interest groups all try to influence, often for good moral reasons. If a government institutes a campaign to limit smoking by displaying gruesome pictures of cancer patients on cigarette packets, then this has the explicit intention of dissuading people from smoking without ostensibly interfering with their basic right to freedom. We mentioned the idea of nudging in Chap. 3, in the context of

privacy (Benartzi et al. 2017), which constitutes a similar type of intervention. While nudging is contentious, certainly when done by governments, it is not always and fundamentally unethical.

So perhaps the reference to freedom or liberty as the cause of ethical concerns in the case of manipulation is not fruitful in the discussion of the Cambridge Analytica case. A related alternative that is well established as a mid-level principle from biomedical ethics (Childress and Beauchamp 1979) is that of autonomy. Given that biomedical principles including autonomy have been widely adopted in the AI ethics debate, this may be a more promising starting point. Respect for autonomy is, for example, one of the four ethical principles that the EU's High-Level Expert Group bases its ethics guidelines for trustworthy AI on (AI HLEG 2019). The definition of this principle makes explicit reference to the ability to partake in the democratic process and states that "AI systems should not unjustifiably subordinate, coerce, deceive, manipulate, condition or herd humans" This suggests that manipulation is detrimental to autonomy as it reduces "meaningful opportunity for human choice" (ibid: 12).

This position supports the contention that the problem with manipulation is its detrimental influence on autonomy. A list of requirements for trustworthy AI starts with "human agency and oversight" (ibid: 15). This requirement includes the statement that human autonomy may be threatened when AI systems are "deployed to shape and influence human behaviour through mechanisms that may be difficult to detect, since they may harness sub-conscious processes" (ibid: 17). The core of the problem, then, is that people are not aware of the influence that they are subjected to, rather than the fact that their decisions or actions are influenced in a particular way.

This allows an interesting question to be raised about the first case study (Facebook and Cambridge Analytica). Those targeted were not aware that their data had been harvested from Facebook, but they may have been aware that they were being subjected to attempts to sway their political opinion – or conceivably might have been, if they had read the terms and conditions of Facebook and third-party apps they were using. In this interpretation the problem of manipulation has a close connection to the question of informed consent, a problem that has been highlighted with regard to possible manipulation of Facebook users prior to the Cambridge Analytica event (Flick 2016).

The second case (pushing sales during "prime vulnerability moments") therefore presents an even stronger example of manipulation, because the individuals subjected to AI-enabled interventions may not have been aware of this at all. A key challenge, then, is that technology may be used to fundamentally alter the space of perceived available options, thereby clearly violating autonomy.

Coeckelbergh (2019) uses the metaphor of the theatre, with a director who sets the stage and thereby determines what options are possible in a play. AI can similarly be used to reveal or hide possible options for people in the real world. In this case manipulation would be undetectable by the people who are manipulated, precisely because they do not know that they have further options. It is not always possible to fully answer the question: when does an acceptable attempt to influence someone

turn into an unacceptable case of manipulation? But it does point to possible ways of addressing the problem.

## 5.4 Responses to Manipulation

An ethical evaluation of manipulation is of crucial importance in determining which interventions may be suitable to ensure that AI use is acceptable. If the core of the problem is that political processes are disrupted and power dynamics are affected in an unacceptable manner, then the response could be sought at the political level. This may call for changes to electoral systems or maybe the breaking up of inappropriately powerful large tech companies that threaten existing power balances, as proposed by the US senator and former presidential candidate Warren (2019) and others (Yglesias 2019). Similarly, if the core of the ethical concern is the breach of data protection and privacy, then strengthening or enforcing data protection rules is likely to be the way forward.

While such interventions may be called for, the uniqueness of the ethical issue of manipulation seems to reside in the hidden way in which people are influenced. There are various ways in which this could be addressed. On one hand, one could outlaw certain uses of personal data, for example its use for political persuasion. As political persuasion is neither immoral in principle nor illegal, such an attempt to regulate the use of personal data would likely meet justified resistance and be difficult to define and enforce legally.

A more promising approach would be to increase the transparency of data use. If citizens and consumers understood better how AI technologies are used to shape their views, decisions and actions, they would be in a better position to consciously agree or disagree with these interventions, thereby removing the ethical challenge of manipulation.

Creating such transparency would require work at several levels. At all of these levels, there is the need to understand and explain how AI systems work. Machine learning is currently the most prominent AI application that has given rise to much of the ethical discussion of AI. One of the characteristics of machine learning approaches using neural networks and deep learning (Bengio et al. 2021) is the opacity of the resulting model. A research stream on explainable AI has developed in response to this problem of technical opacity. While it remains a matter of debate whether explainability will benefit AI, or to what degree the internal states of an AI system can be subject to explanation (Gunning et al. 2019), much technical work has been undertaken to provide ways in which humans can make sense of AI and AI outputs. For instance, there have been contributions to the debate highlighting the need for humans to be able to relate to it (Miller 2019; Mittelstadt et al. 2019). Such work could, for example, make it clear to individual voters why they have been selected as targets for a specific political message, or to consumers why they are deemed to be suitable potential customers for a particular product or service.

Technical explainability will not suffice to address the problem. The ubiquity of AI applications means that individuals, even if highly technology-savvy, will not have the time and resources to follow up all AI decisions that affect them and even less to intervene, should these be wrong or inappropriate. There will thus need to be a social and political side to transparency and explainability. This can include the inclusion of stakeholders in the design, development and implementation of AI, which is an intention that one can see in various political AI strategies (Presidency of the Council of the EU 2020; HM Government 2021).

Stakeholder involvement is likely to address some of the problems of opacity, but it is not without problems, as it poses the perennial question: who should have a seat at the table (Borenstein et al. 2021)? It will therefore need to be supplemented with processes that allow for the promotion of meaningful transparency. This requires the creation of conditions where adversarial transparency is possible, for instance where critical civil society groups such as Privacy International[1] are given access to AI systems in order to scrutinise those systems as well as their uses and social consequences. To be successful, this type of social transparency will need a suitable regulatory environment. This may include direct legislation that would force organisations to share data about their systems; a specific regulator with the power to grant access to systems or undertake independent scrutiny; and/or novel standards or processes, such as AI impact assessments, whose findings are required to be published (see Sect. 2.4.1).

## 5.5 Key Insights

This chapter has shown that concerns about manipulation as an ethical problem arising from AI are closely related to other ethical concerns. Manipulation is directly connected to data protection and privacy. It has links to broader societal structures and the justice of our socio-economic systems and thus relates to the problem of surveillance capitalism. By manipulating humans, AI can reduce their autonomy.

The ethical issue of manipulation can therefore best be seen using the systems-theoretical lens proposed by Stahl (2021, 2022). Manipulation is not a unique feature that arises from particular uses of a specific AI technology; it is a pervasive capability of the AI ecosystem(s). Consequently what is called for is not one particular solution, but rather the array of approaches discussed in this book. In the present chapter we have focused on transparency and explainable AI as key aspects of a successful mitigation strategy. However, these need to be embedded in a larger regulatory framework and are likely to draw on other mitigation proposals ranging from standardisation to ethics-by-design methodologies.

[1] https://privacyinternational.org/.

## References

AI HLEG (2019) Ethics guidelines for trustworthy AI. High-level expert group on artificial intelligence. European Commission, Brussels. https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419. Accessed 25 Sept 2020

Benartzi S, Besears J, Mlikman K et al (2017) Governments are trying to nudge us into better behavior. Is it working? The Washington Post, 11 Aug. https://www.washingtonpost.com/news/wonk/wp/2017/08/11/governments-are-trying-to-nudge-us-into-better-behavior-is-it-working/. Accessed 1 May 2022

Bengio Y, Lecun Y, Hinton G (2021) Deep learning for AI. Commun ACM 64:58–65. https://doi.org/10.1145/3448250

Berlin I (2002) Liberty. Oxford University Press, Oxford

Borenstein J, Grodzinsky FS, Howard A et al (2021) AI ethics: a long history and a recent burst of attention. Computer 54:96–102. https://doi.org/10.1109/MC.2020.3034950

Cadwalladr C, Graham-Harrison E (2018) How Cambridge analytica turned Facebook 'likes' into a lucrative political tool. The Guardian, 17 Mar. https://www.theguardian.com/technology/2018/mar/17/facebook-cambridge-analytica-kogan-data-algorithm. Accessed 1 May 2022

Cambridge Dictionary (n.d.) Manipulation. https://dictionary.cambridge.org/dictionary/english/manipulation. Accessed 11 May 2022

Childress JF, Beauchamp TL (1979) Principles of biomedical ethics. Oxford University Press, New York

Coeckelbergh M (2019) Technology, narrative and performance in the social theatre. In: Kreps D (ed) Understanding digital events: Bergson, Whitehead, and the experience of the digital, 1st edn. Routledge, New York, pp 13–27

European Parliament (2020) The ethics of artificial intelligence: issues and initiatives. European Parliamentary Research Service, Brussels. https://www.europarl.europa.eu/RegData/etudes/STUD/2020/634452/EPRS_STU(2020)634452_EN.pdf. Accessed 1 May 2022

Flick C (2016) Informed consent and the Facebook emotional manipulation study. Res Ethics 12. https://doi.org/10.1177/1747016115599568

Gunning D, Stefik M, Choi J et al (2019) XAI: explainable artificial intelligence. Sci Robot 4(37). https://doi.org/10.1126/scirobotics.aay7120

HM Government (2021) National AI strategy. Office for Artificial Intelligence, London. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1020402/National_AI_Strategy_-_PDF_version.pdf

Kant I (1998) Groundwork of the metaphysics of morals. Cambridge University Press, Cambridge

Maynard A (2019) AI and the art of manipulation. Medium, 18 Nov. https://medium.com/edge-of-innovation/ai-and-the-art-of-manipulation-3834026017d5. Accessed 15 May 2022

Mill JS (1859) On liberty and other essays. Kindle edition, 2010. Digireads.com

Miller T (2019) Explanation in artificial intelligence: insights from the social sciences. Artif Intell 267:1–38. https://doi.org/10.1016/j.artint.2018.07.007

Mind Matters (2018) AI social media could totally manipulate you, 26 Nov. https://mindmatters.ai/2018/11/ai-social-media-could-totally-manipulate-you/. Accessed 15 May 2022

Mittelstadt B, Russell C, Wachter S (2019) Explaining explanations in AI. In: Proceedings of the conference on fairness, accountability, and transparency (FAT*'19). Association for Computing Machinery, New York, pp 279–288. https://doi.org/10.1145/3287560.3287574

Muller C (2020) The impact of artificial intelligence on human rights, democracy and the rule of law. Ad Hoc Committee on Artificial Intelligence (CAHAI), Council of Europe, Strasbourg. https://rm.coe.int/cahai-2020-06-fin-c-muller-the-impact-of-ai-on-human-rights-democracy-/16809ed6da. Accessed 2 May 2022

PHD Media (2013) New beauty study reveals days, times and occasions when U.S. women feel least attractive, 2 Oct. https://www.prnewswire.com/news-releases/new-beauty-study-reveals-days-times-and-occasions-when-us-women-feel-least-attractive-226131921.html. Accessed 11 May 2022

Polonski V (2017) The good, the bad and the ugly uses of machine learning in election campaigns, 30 Aug. Centre for Public Impact, London. https://www.centreforpublicimpact.org/insights/good-bad-ugly-uses-machine-learning-election-campaigns. Accessed 11 May 2022

Presidency of the Council of the EU (2020) Presidency conclusions: the Charter of Fundamental Rights in the context of artificial intelligence and digital change. Council of the European Union, Brussels. https://www.consilium.europa.eu/media/46496/st11481-en20.pdf. Accessed 1 May 2022

Rosen RJ (2013) Is this the grossest advertising strategy of all time? The Atlantic, 3 Oct. https://www.theatlantic.com/technology/archive/2013/10/is-this-the-grossest-advertising-strategy-of-all-time/280242/. Accessed 11 May 2022

Rosenberg M, Confessore N, Cadwalladr C (2018) How Trump consultants exploited the Facebook data of millions. The New York Times, 17 Mar. https://www.nytimes.com/2018/03/17/us/politics/cambridge-analytica-trump-campaign.html. Accessed 11 May 2022

Stahl BC (2021) From computer ethics and the ethics of AI towards an ethics of digital ecosystems. AI Ethics. https://doi.org/10.1007/s43681-021-00080-1

Stahl BC (2022) Responsible innovation ecosystems: ethical implications of the application of the ecosystem concept to artificial intelligence. Int J Inf Manage 62:102441. https://doi.org/10.1016/j.ijinfomgt.2021.102441

Warren E (2019) Here's how we can break up Big Tech. Medium, 8 Mar. https://medium.com/@teamwarren/heres-how-we-can-break-up-big-tech-9ad9e0da324c. Accessed 15 May 2022

Whittle J (2021) AI can now learn to manipulate human behaviour. The Conversation, 11 Feb. https://theconversation.com/ai-can-now-learn-to-manipulate-human-behaviour-155031. Accessed 15 May 2022

Yearsley Y (2017) We need to talk about the power of AI to manipulate humans. MIT Technology Review, 5 June. https://www.technologyreview.com/2017/06/05/105817/we-need-to-talk-about-the-power-of-ai-to-manipulate-humans/. Accessed 15 May 2022

Yglesias M (2019) The push to break up Big Tech, explained. Vox-Recode, 3 May. https://www.vox.com/recode/2019/5/3/18520703/big-tech-break-up-explained. Accessed 15 May 2022