AI and Data-driven Targeting

Karina Vold, Jessica Whittlestone, Anunya Bahanda, Stephen Cave
Leverhulme Centre for the Future of Intelligence, University of Cambridge

Technological advances are giving us information about people on a more granular level than ever before. Governments and companies can now model and predict the beliefs, preferences, and behaviour of small groups and even individuals - allowing them to "target" interventions, messages, and services much more narrowly.

These new forms of targeting present huge opportunities to make valuable interventions more effective, for example by delivering public services to those most in need of them. However, the use of more fine-grained information about individuals and groups also raises huge risks, challenging key notions of privacy, fairness, and autonomy. Some cases of targeting will be clearly beneficial, and others clearly manipulative, but there will also be a large grey area in between.

Why is this a particular concern now?

Government and companies have long held data about the public, yet this is raising new concerns as a result of a few key advances in how data is collected, processed, and used. One outcome of digitization is that it is now possible to collect vastly *more* and vastly more *personalised* data about individuals' lives: particularly given our widespread use of smartphones, social media, wearables and home hubs. Advances in machine learning also allow us to process much larger amounts of information, and to *draw inferences* from data that go far beyond what is explicitly contained within it. For example, data from "smart home" devices such as how much electricity is being used, or when and what kinds of TV are being watched, could easily be used to draw conclusions about a person's daily habits, preferences, and even personality. Finally, various forms of technology make some forms of targeting easier than ever before: smart devices, applications, and social media make it possible to deliver frequent messages and interventions, sometimes without people even being aware of it.

These advances in turn change the forms of influence that are possible. Policymakers and marketers have been using insights from behavioural science for years to "nudge" people's behaviour. This has been done by making subtle changes to the way choices and information are presented: putting healthy foods at eye level, placing health warnings on cigarette packages, or making organ donation an opt-out choice rather than an opt-in.[1] However, the

[1] Leonard, Thomas C. "Richard H. Thaler, Cass R. Sunstein, Nudge: Improving decisions about health, wealth, and happiness." (2008): 356-360.

technological advances described above mean that these changes to "choice architecture" may no longer be the same for everyone. Instead, the information or options a person sees may be tailored to them based on assumptions about the demographic they belong to, or even based on assumptions about their individual characteristics. Traditional attempts to "nudge" behaviour have faced ethical challenges, but have mostly been "one size fits all", based on general theories about human behaviour, not specific individuals or groups. This is no longer the case.

Balancing the opportunities and risks

Targeting presents many new opportunities, including the ability to **deliver better public services to those most in need**, for example by identifying vulnerable, isolated populations and providing them with better access to social services.[2][3] It can help us make **more efficient use of resources**, for example by identifying which groups are most likely to get into road traffic accidents, and targeting interventions specifically at those groups.[4] Even the simple ability to **save people time and energy** may have large benefits, as we're able to provide people with increasingly tailored recommendations for what to watch, read, or buy. Targeting also **doesn't necessarily just apply to individuals**: a more data-driven understanding of companies could also enable government to track and influence their behaviour in ways that benefit society.

However, along with these new opportunities come risks. One serious concern is that targeting methods might be used by self-interested parties to **manipulate people in harmful ways**. A powerful company might use social media data, for example, to identify people suffering from anxiety or depression, and then explicitly target advertisements or messages designed to exploit those vulnerabilities for commercial or political benefit.

What makes some cases particularly worrying is the **objective** of the targeting: a company might be trying to *change* people's preferences or behaviour - to get people to vote for a given political party, say. Contrast this with a case where targeting instead aims to better serve *existing* preferences, such as tailoring music recommendations based on your listening history.[5] A useful question to ask here might be, "to what extent are the interests of the 'targeter' aligned with those of the person being targeted?" A second factor raised here is **the level of transparency**: most people are aware that advertisers are giving them personalised content,

[2] Cordell, Katharan D., and Lonnie R. Snowden. "Population Targeting amid Complex Mental Health Programming: Are California's Full Service Partnerships Reaching Underserved Children?" American Journal of Orthopsychiatry 87, no. 4 (2017): 384-91. doi:10.1037/ort0000194.
[3] Newall, Nancy EG, and Verena H. Menec. "Targeting socially isolated older adults: a process evaluation of the senior centre without walls social and educational program." Journal of Applied Gerontology 34, no. 8 (2015): 958-976.
[4] Sanders Michael, Lawrence James, Gibbons Dan, Calcraft Paul "Using Data Science in Policy" The Behavioural Insights Team. December 14, 2017. Available online: http://38r8om2xjhhl25mw24492dir-wpengine.netdna-ssl.com/wp-content/uploads/2017/12/BIT_DATA-SCIENCE_WEB-READY.pdf.
[5] Of course, what counts as changing preferences will sometimes be unclear: Spotify might gently change preferences through recommendations over time, and some advertisers might argue they are simply providing information.

but may well assume they are seeing the same political campaigns or news articles as everyone else, or be unaware that their personal data has been used to create the targeted content.

Targeting doesn't need to be malicious to raise ethical issues. Even the most well-intentioned attempts to tailor services or information **may seriously threaten individual autonomy**: if individuals only see a narrow subset of the information and choices available, this compromises their ability to explore every option, and therefore to choose freely. One question we might ask here is: what would the **default option** be, in the absence of any targeting? In many cases, it is impossible to present someone with *all* possible information or choices without overwhelming them, and so focusing on what is most relevant does not have to restrict autonomy.[6] However, we might still be concerned that people are seeing a distorted or biased selection, based on their membership in a specific group (e.g. gender, age, or ethnicity), which they never chose to belong to. This raises challenges about not just how *much* it is reasonable to restrict individuals' information or choices, but in what *ways* it is acceptable to do so: what subgroups or features is it ethically acceptable to use as a basis for tailored interventions?

This leads us to the point that targeting **may also threaten important notions of fairness in society**: sometimes, more information makes it easier for us to discriminate against individuals or groups in ways that are harmful. Many of our welfare and social support systems rely on uncertainty about who will lose a job or contract a disease. More information about these things could lead to much more personalised forms of insurance, creating new forms of inequality in society or exacerbating existing ones. More careful thought is needed about what new forms of discrimination - both good and bad - are being enabled by technology, and the impact this could have on society.

Relatedly, we need to be careful about **what assumptions we make about an individual based on their group membership**. Using age to target information about birth control, for example, we might provide women in their 30s with warnings about the side effects of contraception on their fertility, while leaving this out for women in their 20s. The assumption that women in their 30s and not their 20s are thinking about motherhood could be harmful to both groups if wrong: older women may feel pressured, and younger women may lack important information.

More broadly, we need to ask how to draw the line between beneficial and harmful forms of targeting. We have begun to highlight some factors that will be important to consider: the **objective** of the targeting, the level of **transparency**, what the **default** option would be, and **which characteristics and assumptions** are being used to drive targeting, but much more work is needed here.

[6] "It is important to see that autonomy does not require choices everywhere... If we had to make choices about everything that affects us, we would quickly be overwhelmed." Sunstein, Cass R. "The ethics of nudging." Yale J. on Reg. 32 (2015): 413.

The role of the Centre

We think that the Centre for Data Ethics and Innovation (CDEI) can play a crucial role in addressing these challenges in at least two key ways:

1. Identifying areas where targeting is particularly likely to be beneficial

We have outlined a few ways in which targeting could be used to improve lives, but many more could be found. The CDEI is in a unique position to more systematically review a wide range of policy areas alongside domain experts, to identify areas where data-informed targeting could improve the delivery of important policies. Where could services be much more effective if we better understood who was most in need of them, or were better able to match specific interventions to individuals?

Here we strongly advocate a **"policy first, data second"** approach. It can be tempting to start by asking, "what data do I have?", but this can be ineffective and sometimes even harmful: if the way our current data segments a population reflects historical bias or prejudice, for example. By starting with a clear case for the benefit, and then working backwards to collect data on which groups to target, it is much easier to ensure that targeting is not being used to exploit or manipulate individuals.

2. Setting ethical standards, drawing on expertise and public opinion

Since many cases of targeting may not be ethically clear-cut, it will be important going forwards to have a set of standards to guide when and how it is acceptable to target interventions and services narrowly. The setting of standards should crucially draw on an understanding of public opinion about the acceptability of various forms of targeting, as well as various forms of expertise.

Useful research has been done to understand the public acceptability of different types of interventions and "nudging", especially in the health sector.[7] Similarly, academic experts have been thinking hard about how to ensure that government intervention, in general, does not undermine important ethical values such as fairness, autonomy, and dignity.[8] However, this research needs to be extended to the kinds of more targeted intervention that data and AI now make possible. By supporting research and public engagement on these topics, the CDEI could gain valuable insights into how to craft targeted policies that both respect important ethical standards and can gain public support.

---

[7] Hollands, Gareth J., Ian Shemilt, Theresa M. Marteau, Susan A. Jebb, Michael P. Kelly, Ryota Nakamura, Marc Suhrcke, David Ogilvie. "Altering choice architecture to change population health behaviour: a large- scale conceptual and empirical scoping review of interventions within micro-environments." *BMC Public Health*, 13, 1218 (2013).

[8] Sunstein, Cass. "The ethics of nudging." Yale Journal on Regulation, 32, 2 (2015).