UNIVERSITY OF CAMBRIDGE
CFI LEVERHULME CENTRE FOR THE FUTURE OF INTELLIGENCE

The potential harms of online targeting

Response to the call for evidence on targeting from the Centre for Data Ethics and Innovation.

Jess Whittlestone
Leverhulme Centre for the Future of Intelligence
University of Cambridge
jlw84@cam.ac.uk

Karina Vold
Leverhulme Centre for the Future of Intelligence
University of Cambridge
kvv22@cam.ac.uk

Anna Alexandrova
Leverhulme Centre for the Future of Intelligence, Department of History and Philosophy of Science
University of Cambridge
aa686@cam.ac.uk

The authors of this submission are responding in a personal capacity, are happy to be contacted and for this submission to be shared.

We also thank Shahar Avin, Stephen Cave, Jennifer Cobbe, Vesselin Popov and David Stilwell for contributions to and feedback on this submission.

Despite increasing attention being paid to the potential harms of online targeting over the last year, there is still a lack of clarity over what precisely those harms are. To help address this lack of clarity, this submission focuses on question 1: *What evidence is there about the harms and benefits of online targeting*? This question was discussed at a workshop we held on "The Methodology and Ethics of Targeting" at the Leverhulme Centre for the Future of Intelligence in May 2019 (organized by the authors of this submission, and attended by some members of the CDEI). Our submission summarises some of these discussions and attempts to map out some of the key researchers, groups, and publications we know are working on various harms of targeting. This is not intended to be comprehensive, but we hope will help highlight areas worthy of more attention for the CDEI.

Background: what do we mean by targeting?

In the call for evidence, the CDEI defines targeting as "the customisation of products and services online (including content, services standards and prices) based on data about individual users. Instances of online targeting can include online advertising and personalised social media feeds and recommendations."

This definition captures many different forms of targeting, depending on: (a) *who* is doing the targeting (e.g. companies, governments, campaigners); (b) *what* is being targeted (e.g. adverts,

news stories, 'nudges', public services); (c) what *data* is being used and what *assumptions* about users it is used to base targeting on (e.g. assumptions about preferences, personality, or past behaviour), and (d) what *aims* targeting has (e.g. to change beliefs, behaviour, save time or money).

Though this definition focuses on *online* targeting in particular, we suggest that it is worth also looking at some literature on 'offline' targeting, such as personalised medicine and the targeting of public services more generally, as this may provide useful insights into some of the ethical challenges of targeting, as well as how those challenges have been tackled in the past.

For the purposes of conducting a thorough review, it's worth recognising several other terms which are used in different literatures to refer to practices very similar to, if not identical to, targeting as the CDEI defines it:

- 'Recommending' or 'recommender systems' (e.g. as used by Milano, Taddeo and Floridi, 2019). In some domains, the terms 'targeting' and 'recommending' are actually used to refer to distinct activities (outlined in more detail by Cobbe and Singh, 2019). The key difference here is that 'targeting' requires *active and deliberate* selection of audiences by the 'targeter', whereas 'recommending' involves the *automated* selection of content for specific audiences. We take it that the CDEI definition of 'targeting' is intended to encompass both these practices, and we will use it as such. However, it is worth being aware of this distinction, since it has implications for regulation, and potentially also which ethical issues are most relevant.
- 'Behavioural targeting' or 'behavioural advertising' (Boerman et al., 2017). Cobbe and Singh (2019) define 'behavioural targeting' as when platforms or advertisers actively and deliberately selecting groups for targeting based on behavioural tracking and analysis - this therefore sits in the intersection of 'targeting' and 'recommending' as distinguished above.
- 'Personalisation' is also commonly used to refer to a similar activity to targeting. We take this to be a subset of targeting, since not all targeting will necessarily be at the personal level.

We use 'targeting' here broadly, to encompass all of the above terms.

What evidence is there about the harms of online targeting?

We focus on five broad categories of harm from online targeting: (1) threats to privacy; (2) undermining autonomy; (3) impact on vulnerabilities; (4) discrimination; and (5) undermining societal values (particularly cohesion, democracy, and solidarity). Many of these harms will interact with each other: for example, the way targeting impacts vulnerabilities might be thought of as a particularly concerning case of undermining autonomy; discrimination resulting from targeting may have knock-on effects on social cohesion; and threats to privacy within the context of targeting may also threaten to undermine autonomy.

For each category of harm, we highlight key groups and publications focused on understanding this harm.

1. Threats to privacy

Targeting poses risks to individual privacy since it often makes use of personal data, and/or motivates drawing inferences from data which can then be used to categorise individuals and groups for the purpose of targeting.

- **Silvia, Taddeo and Floridi (2019)** discuss several ways that recommender systems can create privacy risks: from how data is collected and shared; the possibility of data leaks; and from inferences about a person that can be drawn from data (which can be highly personal, even if the data collected is not, and where it is also possible to draw personal inferences about one individual from data about *other* people.)
- **Wachter and Mittelstadt (2018)** more explicitly discuss issues surrounding the inferences that can be drawn from data for the purposes of targeting, and explain how current data protection law may be insufficient to protect users here.
- **Skopek (2015)** discusses fundamental questions about the nature of privacy, privacy losses, and privacy violations. He calls into question the widespread assumption that privacy rights should be interpreted as providing rights against personal inferences. Since much of targeting relies on making inferences about personal traits, this paper would suggest that legally protecting one from such inferences could be difficult.

In a sense, these privacy concerns exist independently of targeting, as personal data may be acquired and predictions about people made on the basis of data for purposes other than targeting. Targeting exacerbates these concerns, however, by motivating the collection and processing of personal data for the purposes of profiling. As highlighted in the next section, we also suggest that targeting may introduce new issues insofar as privacy is instrumental in protecting autonomy, and hence that threats to privacy can contribute to undermining autonomy.

2. Undermining autonomy

Several authors have argued that (at least some forms of) online targeting poses special threats to autonomy. By autonomy we mean something like an individual's ability to reflect on and decide freely about their values, actions, and behaviour, and to act on those choices (Vold and Whittlestone, forthcoming). For example:

- **Susser, Roessler and Nissenbaum (2018)** argue that online targeting and personalisation are often forms of *manipulation*. They discuss what distinguishes online manipulation from other forms of online influence, as well as providing a clear account of the relationship between manipulation and autonomy. Key to their account is the idea that online targeting often subverts individuals' decision-making processes towards the "manipulator's" own ends.
- **Costa and Halpern (2019)** similarly raise concerns about how online targeting undermines autonomy by exploiting known cognitive biases, and particularly by

making it easier for people to express or act on 'impulsive' preferences, undermining more reflective choices and behaviours.

- **Milano, Taddeo and Floridi (2019)** outline several ways that recommender systems can encroach on autonomy. They also highlight the related point that algorithmic profiling may affect individuals' ability to form their own *personal identity*, by "removing the users from the social categories that help mediate their experiences of identity." (p.10)
- **Vold and Whittlestone (forthcoming)** suggest that online targeting makes threats to *privacy* and threats to *autonomy* more closely related than ever before, because it is becoming easier to use personal data to influence people's behaviour in ways that undermine self-governance, or autonomy. **Lanzig (2018)** makes a similar point that 'self-tracking technologies' (which provide personalised feedback to users, often for the end of helping them achieve self-improvement goals) should be seen as undermining not just *informational privacy* but also *decisional privacy*: the right against unwanted interference in our decisions and actions (which is closely related to autonomy).

Overall, there is a strong literature making a *theoretical* case that online targeting undermines autonomy (at least in some cases.) There is less *empirical evidence* on where specifically and how these harms are taking place, however it also remains unclear and somewhat controversial, what form such evidence would take.

3. Impact on vulnerabilities

Susser, Roessler and Nissenbaum (2018) draw a useful distinction between "ontological vulnerabilities" - vulnerabilities that all humans share, which arise because of limits on human cognition, from "contingent vulnerabilities", which arise because of structural conditions or differences between individuals (e.g. vulnerabilities resulting from economic disadvantage or being part of a minority group, or suffering from mental health issues.)

It seems that exploiting ontological vulnerabilities is a large part of what can make targeting manipulative in general (undermining rational decision-making processes), but there are also important concerns about how targeting can be used to exploit more contingent vulnerabilities, i.e. groups or individuals who are especially vulnerable.

The above authors give a concrete example of how advertisers were able to use Facebook's platform to target adverts at teenagers when they are feeling stressed, anxious, overwhelmed or similarly vulnerable.

Costa and Halpern (2019) also discuss how time online in general may impact mental health, pointing to evidence of two potential factors: negative feelings due to hurtful interactions or negative content, and substituting time away from wellbeing-enhancing activities. Insofar as online targeting may increase addictive behaviour online and also make it easier for people to access negative content, targeting may exacerbate these problems.

Groups working with particularly vulnerable populations are also beginning to raise concerns about how targeting may affect those groups. For example, the **What Works Centre for**

Children's Social Care recently commissioned a review to be done by the Public Policy team at the Alan Turing Institute into the ethical issues surrounding the use of machine learning in children's social care, including concerns about the ethics of targeting interventions.

4. Discrimination

There is some evidence that forms of targeting can lead to discrimination.

- **Wachter (2019**) discusses discrimination issues associated with 'affinity profiling' used in behavioural advertising, i.e. where people are targeted not based on explicit sensitive characteristics, but rather on assumed interests or characteristics based on 'affinity' wth some group. Wachter cites evidence from a study by **Sweeney (2013)** which shows that behavioural advertising can reinforce racial stereotypes and other forms of discrimination.
- **Speicher et al. (2018)** outline different ways that Facebook advertisers can target in a discriminatory manner (i.e. including or excluding users from adverts based on sensitive categories like race.) **Jan and Dwoskin (2019)** highlight how Facebook is currently being sued by the US Department of Housing and Urban Development for allowing advertisers to target housing advertisements based on race, gender, and other protected characteristics. There are two key aspects to how this discrimination in targeted advertising is possible: (1) Facebook had explicit targeting options to exclude certain groups from adverts **(Angwin, Tobin, and Varner, 2017)**, but also (2) algorithms for targeting exacerbate the problem because they show adverts to more people who are 'similar to' those who have already clicked on it **(Ali et al., 2019)**
- **The Data Justice Lab** at Cardiff University, co-directed by Dr. Lina Dencik, Dr. Arne Hintz, and Dr. Joanna Redden, are also doing important work on the harms of the "datafication" of society, particularly focusing on discrimination and impacts on social justice - see for example the Data Harm Record[1] which provides a running record of harms that have been caused by 'datafication'.
- **Ribeiro et al. (2019)** draw a useful distinction between targeting *opportunities* (e.g. ads for employment, housing, banking services), and targeting *divisive issues* (e.g. information about immigration, Brexit, same-sex marriage, abortion), suggesting that the two raise distinct concerns. Targeting opportunities is more likely to lead to explicit discrimination, while targeting divisive issues is more likely to create broader societal discord. That said, the two may be more closely related than they seem, since explicit discrimination may lead indirectly to undermining social cohesion via making certain groups feel excluded from society, as suggested by **Costa and Halpern, 2019.**

5. Undermining societal values

Several different groups and authors discuss how targeting may make it harder to trust information and content online, may make propaganda or misinformation campaigns more

[1] https://datajusticelab.org/data-harm-record/

effective, may exacerbate 'filter bubbles' and 'echo chambers', all of these leading to reduced societal cohesion, threatening the public sphere and our ability to cooperate to solve problems:

- **Walker et al. (2019)** raise concerns about how targeting exacerbates the threat of propaganda and misinformation, making it even easier for actors to exploit the intent to persuade and manipulate population at a distance by tailoring messages to specific audiences. **Chessen (2018)** raises similar concerns about how AI will "provide propagandists radically enhanced capabilities to manipulate humans minds." It is worth noting here that the concern is not just that people will be more likely to believe false things, but that widespread dis/misinformation will lead people to question the notion of truth more broadly, and result in widespread mistrust of all information (see for example, this talk on **"The Future of the Post-Truth World"** by **Prof Rae Langton**.)
- Even beyond explicit dis/misinformation, targeting may cause or exacerbate what Data & Society's Danah Boyd calls the "fragmentation of truth" (**Boyd, 2019)**: a world where different groups of people increasingly live in their own very separate realities with different versions of the truth. Similar concerns have been raised by others that targeting will exacerbate filter bubbles and echo chambers **(Costa and Halpern, 2019)**; **Chakraborty et al. (2017)** attempt to clearly define and quantify concerns about such "knowledge segregation".
- The concern here is not *just* that targeting could worsen the fragmentation of society, but that it is likely to also exacerbate conflict and discord between different social groups with competing versions of reality: **Ribeiro et al. (2018)** discuss the impact of targeted political adverts focused explicitly on polarizing topics, for example. The authors present empirical work indicating that targeted adverts on divisive issues are less likely to get reported than non-targeted adverts, as the communities they are targeted at are typically more likely to agree with the content and hence less likely to report their content as problematic.
- We suggest that it is worth the CDEI drawing on evidence and research from those working on misinformation and the fragmentation of truth more generally, in order to think clearly about how targeting interacts with some of these problems: for example, **RAND** work on "truth decay"[2], work from the **Oxford Internet Institute** on "computational propaganda"[3], and work from **King's College London** on "weaponising news" (Ramsey and Robershaw, 2019) .

There is also research on how personalised targeting, especially in public services and opportunities, may undermine ideals of solidarity and citizenship in society.

- There is a substantial literature on solidarity in personalised medicine (see e.g. **Prainsack, 2014, 2018**), which argues that personalisation in healthcare will lead to "greater individualisation of medicine", decreasing people's willingness to contribute to systems that support everyone equally. Though this literature focuses mostly on an 'offline' form of personalisation, it is worth considering how online targeting and personalisation may risk undermining ideals of solidarity that in turn undermine social support systems in other domains (such as e.g. insurance).

[2] https://www.rand.org/research/projects/truth-decay.html
[3] https://www.oii.ox.ac.uk/research/projects/computational-propaganda/

- Hintz, Dencik and Wahl-Jorgensen (2018) **argue that the way citizens are increasingly** "monitored, categorized, sorted and profile" requires us to fundamentally rethink our understanding of digital citizenship.

Concluding remarks

Existing research highlights several different ways that online targeting may (and is already beginning) cause harm to individuals and society. Though by no means comprehensive, we have tried to map out some of the key researchers and groups working on different aspects of these harms. Some of this research has begun to point to concrete evidence of how harms are already occurring, such as the ways that targeting is enabling and exacerbating forms of discrimination and threats to privacy. Other research is more speculative, pointing to harms that seem likely to occur in future given evidence about how targeting is being used across society, such as undermining autonomy, social cohesion, and solidarity. Though the evidence on these harms is currently less solid, we suggest that these more speculative areas may be particularly important for the CDEI to focus on understanding going forwards, precisely because these harms are more likely to go unnoticed (whereas, for example, threats to privacy and discrimination are already receiving more direct attention).

Finally, we should note that of course we recognise the many ways targeting can be beneficial: in providing people with online services much better tailored to their needs, for example. We have chosen to focus on the harms in this submission for the sake of focus and because we feel that they are currently less well understood. We do believe, however, that any policy recommendations related to online targeting must take into account both the harms and benefits and work to balance any tensions between the two.

References

Ali, M., Sapiezynski, P., Bogen, M., Korolova, A., Mislove, A., & Rieke, A. (2019). Discrimination through optimization: How Facebook's ad delivery can lead to skewed outcomes. arXiv preprint arXiv:1904.02095.

Angwin, J. Tobin, A. and Varner, M. (2017). Facebook (Still) Letting Housing Advertisers Exclude Users by Race. ProPublica.

Boerman, S. C., Kruikemeier, S., & Zuiderveen Borgesius, F. J. (2017). Online behavioral advertising: A literature review and research agenda. Journal of Advertising, 46(3), 363-376.

Boyd, D. (2019). The Fragmentation of Truth. Medium.

Cobbe, J., & Singh, J. (2019). Regulating Recommending: Motivations, Considerations, and Principles. Considerations, and Principles (April 15, 2019).

Costa, E. and Halpern, D. (2019). The Behavioural Science of Online Harm and Manipulation. Behavioural Insights Team Report.

Chakraborty, A., Ali, M., Ghosh, S., Ganguly, N., & Gummadi, K. P. (2017). On quantifying knowledge segregation in society. arXiv preprint arXiv:1708.00670.

Chessen, M. (2018). The MADCOM Future: How Artificial Intelligence Will Enhance Computational Propaganda, Reprogram Human Culture, and Threaten Democracy… and What can be Done About It. In Artificial Intelligence Safety and Security (pp. 127-144). Chapman and Hall/CRC.

Hintz, A., Dencik, L., & Wahl-Jorgensen, K. (2018). Digital citizenship in a datafied society. John Wiley & Sons.

Jan, T. and Dwoskin, E. (2019). HUD is reviewing Twitter's and Google's ad practices as part of housing discrimination probe. *The Washington Post.*

Lanzing, M. (2018). "Strongly Recommended" Revisiting Decisional Privacy to Judge Hypernudging in Self-Tracking Technologies. Philosophy & Technology, 1-20.

Milano, S., Taddeo, M., & Floridi, L. (2019). Recommender Systems and their Ethical Challenges. Available at SSRN 3378581.

Prainsack, B. (2014). Personhood and solidarity: what kind of personalized medicine do we want?. Personalized Medicine, 11(7), 651-657.

Prainsack, B. (2018). The "we" in the "me" solidarity and health care in the era of personalized medicine. Science, Technology, & Human Values, 43(1), 21-44.

Ramsey, G. and Robertshaw, S. (2019). Weaponising News: RT, Sputnik and Targeted Disinformation. Policy Institute, King's College London.

Ribeiro, F. N., Saha, K., Babaei, M., Henrique, L., Messias, J., Benevenuto, F., ... & Redmiles, E. M. (2019). On microtargeting socially divisive ads: A case study of russia-linked ad campaigns on facebook. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 140-149). ACM.

Skopek, J. M. (2015). Reasonable expectations of anonymity. Va. L. Rev., 101, 691.

Susser, D., Roessler, B., & Nissenbaum, H. (2018). Online Manipulation: Hidden Influences in a Digital World. Available at SSRN 3306006.

Sweeney, L. (2013). Discrimination in online ad delivery. arXiv preprint arXiv:1301.6822.

Vold and Whittlestone (forthcoming). Privacy, autonomy and personalised targeting: rethinking how personal data is used. Forthcoming in IE Report on Data, Privacy and the Individual.

Wachter, S., & Mittelstadt, B. D. (2018). A right to reasonable inferences: re-thinking data protection law in the age of Big Data and AI. Columbia Business Law Review.

Wachter, S. (2019). Affinity Profiling and Discrimination by Association in Online Behavioural Advertising. Available at SSRN.

R Walker, C., Terp, S. J., C Breuer, P., & Crooks, L. (2019, May). Misinfosec. In Companion Proceedings of The 2019 World Wide Web Conference (pp. 1026-1032). ACM.