TWON Policy Brief #1

On the Ethics of Using Twins of Online Social Networks

October 2024

Many warn that online social networks such as Facebook, X (Twitter) or Telegram are contributing to worrying social dynamics such as the polarisation of opinion, the spread of fake news, conspiracy theories, discrimination and large-scale collective outrage. However, demonstrating that online social networks have contributed to the emergence of the undesired outcomes has proven elusive. Scientific reviews of research on the impact of filter bubbles have indeed yielded inconclusive findings, with arguments and evidence supporting both sides of the debate. Tech companies, therefore, find it easy to sidestep all allegations.

There is a way to **overcome this responsibility ping-pong** by creating an analogous technology: Digital Twins of ONline social networks, **TWONs**. These highly advanced and realistic computer models mimic an original online social network as closely as possible. This makes it possible to quantify the extent to which an online social network, as well as specific algorithms, yield undesirable outcomes. Furthermore, they offer a means to optimize the design of online social networks with respect to social, ethical, and epistemic objectives. Accordingly, TWONs might be a **powerful tool for regulating online social networks**.

On the other hand, taming one technology by creating another can give rise to a number of risks of its own. With regard to TWONs, **societal risks** are immediately conceivable: By leveraging vast datasets about users and by intricately representing user behaviour, TWONs have the potential to be used in ways that are detrimental to the interests of individuals and societies alike. This is why we must carefully examine ethical implications of different modes of regulating TWONs, so that decision makers get the tools at hand to make a well-founded decision. There needs to be a public discussion on the usage and regulation of TWONs.

Outcomes in a nutshell

Our ethical analysis of TWONs, based on the currently available outlook on its benefits and risks, has demonstrated that:

1. There is a plenitude of options for regulating the technology of TWONs, ranging from
   a. unlimited public access: analogously to available free web-search engines or publicly accessible LLM-chatbots
   b. to a strictly controlled usage: analogously to the governance of sensible technologies (e.g., CERN, applications in nuclear physics) or of sensible data panels (e.g. SOEP).

2. The risks and benefits of the TWON hinge upon the manner and extent to which access to this technology is regulated.

3. Each mode of governance brings with it distinct societal benefits and risks.

Governance Models

Unlimited Public Access		Approved Researchers Access Only	
Possible Benefits	**Risks**	**Possible Benefits**	**Risks**
– High economic gains	– Intensified undermining of individual autonomy from freely available tools for manipulation, mis- and disinformation	– Control of online social networks	– Turns out to be ineffective for control of online social networks
– Public control of online social networks: prevention or disclosure of manipulation, mis- and disinformation	– Collapse of institutions necessary for democratic governance	– Knowledge gains from better measurement of social realities (though smaller than in free public access)	– Restriction of access to TWONs turns out to be infeasible (the access to a certain group turns out not to be restrictable): hidden undermining of individual autonomy
– Unrestricted knowledge gains from better measurement of social realities	– Reinforcement of existing inequalities in financial and political power		

Given what is now known about the possible consequences of the use of TWON, none of the governance modes discussed in this report – from unrestricted access to access only with the authorization of an authority – can be rejected on sound grounds.

a) We cannot currently rule out the scenario that a free public access to TWONs is the sole means by which deployment of manipulative, mis- or disinformative algorithms on online social networks can be revealed and thereby publicly controlled. If this proves to be the case, this would provide a weighty reason for a free public access to TWONs.

b) Based on current knowledge, it is also possible that restricting access to TWONs to approved researchers is sufficient to control the algorithms of online social networks. This, in turn, would be a weighty reason to restrict TWON's availability to a group of researchers.

Since it is unclear which of these scenarios is more realistic, it is impossible to reject one of the governance modes.

Empirical and Ethical Uncertainties

To enable an informed decision on the use of TWONs, future research and deliberation are needed to resolve uncertainties in the evaluation of different governance modes.

Empirical uncertainties:

a) Feasibility of regulating access to TWONs: It is currently uncertain if it will be practically possible to restrict access to TWONs. This depends on the complexity of underlying technology (if, after the blueprint for the underlying models has been developed, anybody with sufficient financial and/or computational resources can set up TWONs, it is unlikely that access restrictions will become enforcable) and on the amount of personal data from an online social network needed for a reliable simulation.

b) Availability of alternative means for safeguarding democratic values and enforcing legislation on online social networks: the recently adopted legislation in the EU (namely, the Digital Services Act (DSA) and the Artificial Intelligence Act (AIA)) is designed to regulate the activities of large online platforms. Nevertheless, it is currently unclear to what extent these acts can be enforced. It may be the case that an instrument such as a TWON is required for the two acts to become legally effective.

Ethical uncertainties:

a) Quantification of the extent to which norms, values, and rights worthy of protection (such as democratic self-determination, individual autonomy, the right to informational self-determination) are jeopardized by online social networks: the decision regarding the governance of TWONs is contingent upon the extent to which these values are threatened by the prevailing online social networks and the diverse governance modes of TWONs. At the moment, no comparisons of the threats are available.

b) How should the differences in the potential benefits and risks of different modes of governance of TWONs be weighed? Unrestricted public access promises the highest economic benefits from TWONs, but this mode of governance is also associated with the highest risks. The stricter the regulation of TWONs is, the lower are as the expected economic benefits as the risks. Assessments of societal risks and benefits are often highly controversial within a society.

Key Takeaways

1. If TWONs turn out to be a technology that requires strict regulation, the research and development process must also be subject to regulatory oversight.
2. At the moment, it is unclear how much a TWON's ability to inform the regulation of online social networks hinges on detailed personal data about individual users. By means of modeling of fictional reality, however, the reliance on personal data can be rigorously quantified. We recommend conducting such an analysis.
3. Additionally, the report lays out the **methodology** used for the ethical analysis. This methodology – reconstruction and analysis of arguments – allows developers as well as interested stakeholders to **reflect on ethical controversies** in TWON's research and societal governance.