TWin of Online Social Networks

Deliverable D7.2

IMPACT ASSESSMENT AND ETHICAL GUIDANCE HANDBOOK

Main Authors: Eugen Pissarskoi and Michael Mäs

Funded by the European Union

About TWON

TWON (project number 101095095) is a research project, fully funded by the European Union, under the Horizon Europe framework (HORIZON-CL2-2022-DEMOCRACY-01, topic 07). TWON started on 1 April 2023 and will run until 31 March 2026. The project is coordinated by the Universiteit van Amsterdam (the Netherlands) and implemented together with partners from Universität Trier (Germany), Institut Jozef Stefan (Slovenia), FZI Forschungszentrum Informatik (Germany), Karlsruher Institut für Technologie (Germany), Robert Koch Institute (Germany), Univerzitet u Begogradu - Institut za Filozofiju I Drustvenu (Serbia) and Slovenska Tiskovna Agencija (Slovenia).

Funded by the European Union. Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

Project Name	Twin of Online Social Networks
Project Acronym	TWON
Project Number	101095095
Deliverable Number	D7.2
Deliverable Name	Impact assessment and ethical/legal guidance handbook
Due Date	31.03.2024
Submission Date	28.03.2024
Type	R — Document, report
Dissemination Level	PU - Public
Work Package	WP 7
Lead beneficiary	5-KIT
Contributing beneficiaries and associated partners	Universiteit van Amsterdam (UvA), Universität Trier (UT), Institut Jozef Stefan (JSI), FZI Forschungszentrum Informatik (Germany), Karlsruher Institut für Technologie (KIT), Robert Koch Institut (RKI), Univerzitet u Begogradu - Institut za Filozofiju I Drustvenu (UoB), Slovenska Tiskovna Agencija (STA)

Executive Summary

Many caution that online social networks contribute to undesirable social dynamics such as opinion polarization, the spread of fake news, conspiracy theories, discrimination, and large-scale collective outrage. Although these phenomena are well documented in the scientific literature, demonstrating that online social networks have contributed to their emergence has proven elusive. Digital twins of online social networks, TWONs, hold the promise of addressing this problem. These highly advanced and realistic computer models enable the quantification of the extent to which online social networks, as well as specific algorithm design choices, 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 means to regulate the design of online social networks.

In this handbook, we delve into the ethical dimensions of TWONs. Leveraging vast datasets about users and intricately representing user behavior, TWONs have the potential to be wielded against the interests of individuals and societies alike. Our intent is to **catalyze a discourse** surrounding the potential hazards of TWONs and strategies for mitigating them. We aim to equip TWON **developers** as well as **decision-makers** with the arguments needed to navigate this technology responsibly, mindful of its implications and poised to harness its capabilities without engendering adverse consequences. We **identify open research questions** that empirical researchers as well as modellers should address to inform a public debate and allow decision-makers forming an informed decision about the deployment of TWONs.

This report lays out a **methodology** allowing developers as well as interested stakeholders to **identify ethical controversies** in TWON's research and societal governance. These controversies depend on the available state of foreknowledge about benefits and risks of TWONs. This foreknowledge changes – it grows – in the course of technology's development. The methodology as described in this report provides a blueprint for updating the ethical analysis in the light of the changing empirical knowledge.

Our **analysis of the ethical controversies** over development and use of TWONs based on the currently available outlook on its benefits and risks has demonstrated:

Insight 1 *The risks and benefits of the TWON hinge upon the manner and extent to which access to this technology is regulated. There is a plenitude of options, ranging from unlimited access to a very strictly controlled usage. Each mode of governance entails distinct societal benefits and risks.*

Insight 2 *In the light of what is now known about the possible consequences of the use of TWON, no governance mode discussed in this report – from unrestricted access to access only with the authorization*

by a judicial authority – can be rejected on sound grounds. This result does not imply that all governance modes are morally acceptable.

Insight 3 *Should new information emerge, regulating the usage of TWONs could prove to be necessary.*

Insight 4 *Argument analysis has revealed premises in the justifications of different governance modes that we deem uncertain. These premises need to be addressed by future research or thorough deliberation before an informed decision on the deployment of TWONs is possible.*

We also derived conclusions about the process of developing a TWON. In a nutshell, we argue:

Insight 5 *If TWONs turn out to be a technology that requires strict regulation, the research and development process must also be subject to regulatory oversight.*

Insight 6 *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.*

Contents

List of Tables

List of Figures

List of Abbreviations

DSA Digital Services Act

F-OSN Fictional Online Social Network

FVO Fictional Virtual Object

GM Governance Mode

OSN Online Social Network

TA Technology Assessment

TWON Twin of an Online Social Network

Impact Assessment and Ethical Guidance Handbook

Eugen Pissarskoi and Michael Mäs*

March 27, 2024

1 Introduction

Experts, researchers and political decision-makers caution that online social networks (OSNs) have precipitated detrimental shifts in public discourse. OSNs have been blamed for disseminating misinformation, enabling foreign interference in elections, and radicalizing users, culminating in instances of riots and violent protests. For instance, it has been argued that personalization algorithms governing users' information diets foster so-called filter bubbles and echo chambers, wherein users' viewpoints are reinforced, exacerbating the polarization of political opinions (Pariser, 2011; Keijzer and Mäs, 2022). This apprehension is widespread, echoed by public figures such as Barack Obama, who warned that many "retreat into our own bubbles [...] especially our social media feeds, surrounded by people who look like us and share the same political outlook and never challenge our assumptions" (Obama, 2017). Germany's President, Frank-Walter Steinmeier, has gone so far as to attribute political unrest and societal fragmentation to the proliferation of filter bubbles (Steinmeier, 2017).

Tech companies, however, find it easy to sidestep these allegations. When asked why he refused to "at least admit that Facebook played a central role or a leading role in facilitating the recruitment, planning, and execution of the attack on the Capitol", Zuckerberg, the CEO of Facebook, pointed to "the people who spread that content, including the President but others as well, with repeated rhetoric over time saying that the election was rigged and encouraging people to organize. I think that those people bear the primary responsibility as well." (House of Representatives, 2021). As a matter of fact, it is hard to counter Zuckerberg's argumentation. Scientific reviews of research on the impact of filter bubbles have indeed yielded inconclusive findings, with arguments and evidence supporting both sides of the

*We would like to thank Sjoerd Stolwijk and Alenka Guček for reviewing the internal draft of the paper. The report has benefited considerably from the comments.

debate (Zhuravskaya et al., 2020; Bruns, 2019; Keijzer and Mäs, 2022). Correspondingly, proposals for regulation of OSNs remain contested as well (Persily and Tucker, 2020). The problem is very fundamental. In order to demonstrate that online social networks or specific algorithms installed on them have deleterious effects, one needs to compare our societies with a world without these communication systems or systems controlled by different algorithms. This counter-factual comparison, obviously, does not exist, making it very difficult to unequivocally confirm or refute any responsibility of online social networks. Digital twins of online social networks (TWONs), however, promise to provide a solution to this fundamental problem.

1.1 TWONs - Digital Twins of online social networks

Digital twins are computer models of real complex systems that represent these systems with such precision that the model is deemed a ”twin” (c.f. Rasheed et al. 2020; Wright and Davidson 2020; Barn 2022). Digital twins proof to be a powerful tool in various contexts. NASA, for instance, employs digital twins of space vehicles since it is often impossible to directly investigate these systems when they are in space. Additionally, also physical replicas of a space vehicle left on Earth are often not informative since they are not exposed to the harsh environment of space. A comprehensive computer model can simulate external forces, aiding in identifying malfunctions' root causes. Likewise, in supply chain management, digital twins are pivotal for optimizing operations. They replicate the entire supply network, forecast potential disruptions, and enhance overall efficiency. In the automotive sector, digital twins play a crucial role in crash testing, facilitating the creation of virtual vehicle replicas for simulating and evaluating safety measures. This reduces reliance on physical prototypes.

To achieve the necessary realism of digital twins, these computer models are usually fed with detailed empirical data. In fact, digital twins are constantly updated with real-time information about external forces, the state of the system and its components. The result is a highly complicated formal model that can be described as a black-box but that can be considered a reliable prediction-machine capturing all relevant external and internal processes of the system it represents.

TWONs are digital twins of online social networks. These models consist of two main ingredients. First, there is a very detailed and realistic description of the network's *users*, modelling all relevant user behavior and user characteristics. These user models are either based on theories of human behavior translated into computer code (Flache et al., 2017) or they use AI such as large-language models to mimic the behavior of users (Betz, 2022). Second, the TWON has a ”platform model” that represents the structure users are interacting on. This platform model describes the affordances and restrictions users experience as well as any algorithm influencing the content users are exposed to.

The degree to which a digital replica represents its original object varies depending on the target system. Whereas digital twins in engineering might come close to digital copies of the physical object (representing all properties which determine the behavior of the target object), digital twins of complex socio-ecological systems (such as urban areas, agricultural systems, oceans, biodiversity or even the whole Earth, c.f. (Bauer et al., 2021)) will contain certain simplifications due to lacking knowledge about all relevant properties or lack of precise data for determination of their behavior. Still, a twin of a complex social system represents the target system precisely enough to describe and predict dynamic behavior of the properties of interest. Accordingly, a TWON is a virtual replication of a virtual system, an online social network, with a degree of representation which allows monitoring and predicting communication outcomes within the target OSN.

1.2 TWON's Potentials

TWONs hold the potential to inform the discourse surrounding the adverse impacts of online social networks through four avenues. Firstly, TWONs enable rigorous quantification of the repercussions of platform design choices. By comparing TWON realizations with and without specific alterations to the platform model, researchers can pinpoint the effects of these changes. For instance, if modifying a personalization algorithm results in reduced opinion polarization in the TWON, it suggests that this algorithm contributes to polarization. That is, with a TWON it is possible to generate the counter-factual systems needed to demonstrate that the real has specific consequences.

Secondly, TWONs serve as a valuable instrument for developers seeking to optimize OSNs according to economic, social, and ethical principles. They enable experimentation with different design options while quantifying their respective outcomes. Crucially, these experiments can be conducted prior to implementing decisions on the actual platform, mitigating the risk of unforeseen unintended consequences. Tech companies optimize their platforms mainly on economic interests, usually trying to keep users on their platform and to expose them to advertisement for as long as possible. TWONs make it possible for the public to find out how to go beyond economic ideals and to rigorously optimize platforms with social, societal, and ethical principles in mind.

Third, TWONs have the potential to transition the discourse surrounding OSNs from binary yes-no arguments to a nuanced examination of the underlying processes at play on digital communication platforms. The above cited discourse during the Senate hearing illustrates a common pattern. When confronted with criticism, tech companies can dodge allegations and point to other potential causes of undesired effects. If a TWON, in contrast, indicates that a particular design aspect of an OSN yields undesirable effects, it prompts tech companies to scrutinize the specific assumptions embedded within the

model that influence its predictions. While such critique is always feasible, it also compels companies to furnish the data required for testing these model assumptions. This process fosters a constructive dialogue focused on the mechanisms operating within social networks and the necessary research to comprehend the origins of undesirable societal outcomes.

Fourthly, TWONs serve as a tool for conducting rigorous risk assessments. According to Article 34 of the EU's Digital Services Act, "Providers of very large online platforms and of very large online search engines shall diligently identify, analyse and assess any systemic risks in the Union stemming from the design or functioning of their service and its related systems, including algorithmic systems, or from the use made of their services." This risk assessment must encompass "any actual or foreseeable negative effects on civic discourse and electoral processes, and public security". Given that TWONs are grounded in realistic assumptions, their predictions regarding future consequences carry credibility. Indeed, there is no other method that allows for such rigorous assessment of potential adverse effects and their probabilities. Accordingly, TWONs have the potential to play a key role in the future regulation of OSNs, also allowing regulators to anticipate the consequences of restrictions imposed on the design of OSNs.

Summary 1

A TWON, short for "Twin of Online Social Network," is a computer model that replicates the dynamics of a real online social network (OSN) to such an extent that it can be likened to a "twin" of the actual communication platform. It embodies all pertinent characteristics of the network and its users, drawing upon detailed empirical data to ensure realism.
TWONs can be instrumental in various capacities:

- Rigorously quantifying social and societal ramifications of platform design decisions.
- Optimizing OSNs based on economic, social, ethical, and epistemic principles.
- Transitioning the discourse surrounding the effects of OSNs from binary yes-no arguments to a nuanced examination of the underlying processes at play.
- Conducting rigorous risk assessments of OSNs and informing regulatory efforts aimed at shaping the design of digital communication platforms.

1.3 Why an ethical guidance handbook?

Recent strides in modeling opinion dynamics on networks (Flache et al., 2017) in conjunction with expanding computational capabilities and advancements in large-language models, have paved the way for the development of TWONs. However, akin to the technology they aspire to regulate and enhance, TWONs themselves carry the potential for significant adverse impacts on individuals, communities, and societies. In fact, TWONs may inadvertently exacerbate some of the very undesired effects this technology aims to mitigate.

A virtual system capable of monitoring and predicting the behavior of a target system, which in turn influences social dynamics, possesses the inherent potential to impact those dynamics itself. In the case of a digital twin of an online social network, with its ability to monitor and predict communicative dynamics within the network, it can be harnessed for diverse purposes. On one hand, it can aid in identifying mechanisms within the OSN that propagate misinformation or exacerbate polarization. Conversely, it could also be leveraged to pinpoint mechanisms through which misinformation is disseminated more effectively within the OSN. Furthermore, the development of a sufficiently accurate TWON may necessitate use of detailed user data. This information could potentially be exploited against the very individuals the model aims to help safeguard. Additionally, when supplied with user data, a TWON is capable of generating predictions regarding user beliefs, opinions, and behavior. These projections have the potential to encroach upon the fundamental rights of citizens.

Aims of this document Societies must decide on how to deal with the technology once a TWON has been successfully developed in light of the societal benefits which could be realized by a TWON and the risks associated with its use. Let us stress that this decision is unavoidable once a technology has become available. Sometimes it is made without identifying all the available options for the public governance of the respective technology (including the option to abstain from its deployment) and without publicly reflecting on the supporting and opposing reasons regarding these governance options. For instance, in case of information processing technologies, these decisions have been made by some societal agents – companies which have developed the respective technology – either by providing free access to the technology (Google Search, Online Social Networks, Open AI to name just few), or by restricting access by a willingness to pay or another criterion (e.g., invitations according to certain criteria). As a result, societal benefits and costs from technology's use are distributed in a way that can be criticized for good reasons: technology owning companies reap the profits from technology use, societal agents benefit from the technological services but also bear the costs from its detrimental side-effects.

It may be too late to start debate the question of how to organize a TWONs use and how the benefits and costs should be distributed in a fair way amongst stakeholders once a TWON is technologically feasible. With the advent of a TWON blueprint, it's foreseeable that societal agents with sufficient financial and/or political power will use it for their interests independently of what a public deliberation recommends as a fair governance mode. Hence, initiating an ethical analysis of TWONs at the research outset is imperative.

With this report, we will not justify a recommendation on how a society should deal with a TWON. Firstly, currently available foreknowledge about its benefits and harms is uncertain to a high degree: we cannot reliably quantify a range of possible benefits and harms and have no reliable assessment of the probabilities for their occurrence. In that case, a recommendation for a governance mode of a TWON would presuppose controversial normative decision principles under (Knightian) uncertainty (c.f. Hansson (2023)). Secondly, even if the data for a risk-assessment were available, a recommendation on how to deal with a TWON would presuppose evaluations of the societal importance of specific benefits and risks. However, such evaluative attitudes are highly controversial and it is not clear how to find out which of them is more reasonable (c.f. Hansson (2023)).

Fortunately, there is no immediate need for a practical recommendation at the current level of foreknowledge. Rather, there is a need of a procedure clarifying what aspects should be taken into account within the research process such as to avoid that ethically controversial decisions are made unaware. Accordingly, we aim at the following contributions with this report:

- We present a methodology for identification of controversial or uncertain premises when justifying a practical claim about TWONs, i.e. a (positive) recommendation to use a TWON under certain circumstances C and a (negative) recommendation not to use a TWON under C (see Section 2).
- As a case study, we use this methodology to identify controversial ethical presuppositions when justifying different modes of governance of a TWON in light of the currently available foreknowledge about its potential benefits and moral drawbacks (Sections 3 to 6).
- In Section 7, we summarize the outcomes from our case study. Additionally, we derive recommendations for the research process regarding ethically significant concerns which need to be taken into account in the next steps of the TWON research process (Section 7.2) and suggest a monitoring strategy (Section 7.3).
- We understand this report as a "methodological handbook", i.e. a manual for regular ethical assessments within the process of research and development of the TWON. To avoid inappropriate expectations, we would like to stress that this handbook does not guide on ethical evalua-

tions: there are no answers to the question of whether a certain mode of governance of a TWON is morally mandated, legitimate or disallowed. Nor does it contain a legal assessment of TWONs. The European legislators have recently adopted regulations directly affecting regulation of OSNs and the use of technologies like a TWON: the Digital Services Act (DSA) and the Artificial Intelligence Act (AI-Act).

However, this report does provide orientation on where the ethical controversies lie when it comes to a justification of a certain governance mode of a TWON including consideration of the legal rights provided by the DSA.

- This report aims to provide orientation for researchers of the TWON and for the interested public on what we consider as the most fundamental ethical challenges of this technology. Additionally, the research and development process of TWONs will include further ethically significant components: programming of decision-making algorithms, use of generative AI for content generation on virtual platforms, empirical studies collecting data on users behaviour on online social network and validating behaviour of artificial agents. This report does not provide guidance on these specific research components. For that, a good starting point are the recently published "Guideline on the Responsible Use of Generative AI" (European Commission, 2024) and the more general literature reviews on ethics of algorithms (Mittelstadt et al., 2016) and on guidelines of AI (Hagendorff, 2020).

While a TWON is a formalized model that can be described and analyzed with great detail and rigor, the current ethical assessment relies on arguments that may appear vague. Thus far, research lacks precise quantification of the potential consequences of both OSNs and TWONs. In fact, a TWON itself would be essential for conducting such a rigorous analysis, as previously argued. Consequently, our present analysis necessitates discussion of possibilities and potential scenarios. Nonetheless, the systematic methodology outlined in the following section enables us to draw insightful conclusions.

2 Methodology

Shaping technological research and development by introducing reflexivity into the research process is a typical task of the research field of Technology Assessment (TA). A full-fledged TA as suggested, e.g., by Grunwald (2018) is beyond the scope of this report. However, such a comprehensive approach is not necessary for its goal: identification of ethically significant decisions in the course of the development process (before the decisions have been made) together with a procedure that specifies how to deal

with them. To identify ethically significant concerns within the research process, we use the method of argument analysis (Tetens, 2004; Brun and Betz, 2016; Harrell, 2016; Betz, 2020).

2.1 Argument Analysis

Argument analysis exploits one of the most powerful scientific tools, logic, to evaluate statements about an object (here: TWONs). Logic can be used to analyze whether the truth has been transferred from the premises to the conclusion. If the premises logically imply the statement in the conclusion and if the premises are true, then the statement in the conclusion must be true too.

Accordingly, argument analysis consists of two steps. First, you identify arguments that support or reject a conclusion about an object. Each argument is then reconstructed in a way that it is logically valid. Logical validity means that the premises of the argument logically imply the conclusion. Second, all premises in the reconstructed arguments are evaluated according to certain epistemic norms (truth-value or plausibility of premises, consistence of a position).

The outcome is a set of arguments either supporting or challenging the development of TWONs. Stakeholders can decide whether they accept or reject each argument based on whether or not they accept the premises underlying each argument. When all arguments are considered and weighted, an informed decision about the development and use of TWONs is possible. Often, however, it will turn out that it is hard to determine whether one or multiple premises are supported. In this case, one has identified a gap in the argument that requires future research.

Deductively Valid Reconstruction When humans controversially debate in everyday life, they communicate reasons supporting their stances: they bring forward reasons supporting the claims which we hold to be true (or at least plausible or reasonable)[1] and reasons rejecting the claims they believe to be false (or implausible). Argumentation theory specifies this everyday practice in a systematic way. Pro- and counter-reasons are reconstructed in a form of deductively valid arguments.

An argument is a set of statements with a certain structure: one statement is a conclusion (of the argument), the other statements are its premises. The conjunction of the premises logically implies the conclusion. That means that if the premises are true, the conclusion must be true as well. An argument can be represented in the following form, where the line represents a deduction:

[1] There are different normative standards for an epistemic evaluation of a claim ”p”: truth-value (”p” is true or false), plausibility (”p” is plausible (to a certain degree) or implausible), reasonableness (”p” is reasonable or unreasonable), justifiability (”p” is well-justified or ”p” is unjustified). We do not presuppose a specific theory of epistemic evaluation in this report. We merely presuppose that there is a normative standard for epistemic evaluation, whichever it might be. For this reason, we will use different terms for epistemic evaluation in an interchangeable manner.

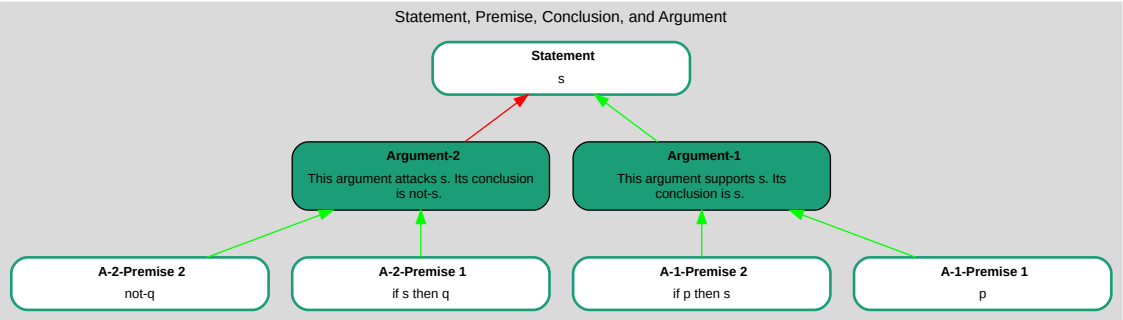

Figure 1: Graphical Representation of Statements and Arguments

A statement ”s” which is supported by an argument and attacked by another argument. Boxes with white background represent statements. Boxes with coloured background represent arguments. Green arrow from a statement to an argument represents that the statement is a premise of this argument. A green arrow from an argument to a statement means that the statement is a conclusion of the argument, or, to put it differently, that the argument supports or justifies the statement. A red arrow from an argument to the statement means that the conclusion of the argument contradicts the statement. In other words: the argument attacks or rejects the statement.
Argument maps have been created with argdown, we follow the representation style and syntax introduced there, see `https://argdown.org/guide/elements-of-an-argument-map.html`.

(1) Premise (1)

(2) Premise (2)

---

(3) Conclusion (3) follows deductively from (1)&(2)

The conclusion of an argument is the statement which the argument supports. If an argument contains an objection to a statement, then the conclusion of the argument is a negation of the statement. Figure 1 shows a visual approache to present arguments, depicting the argumentative relationship in which a statement named ”Statement” with the content ”s” (i.e., ”Statement” claims that s) is justified by an argument <Argument-1> and rejected by another argument <Argument-2>. This means that the conclusion of <Argument-1> is the statement ”s”, whereas the conclusion of <Argument-2> is the negation of s. Each of the two arguments contains at least two premises.

If an argument A_1 attacks another argument A_2, the argumentation theory interprets this as: conclusion of A_1 is the negation of one of the premises of A_2. These dialectical relationships can be represented as an argumentative map depicted in Figure 2. Here, the dialectical structure from Figure 1 is extended by two further arguments[2]. The argument <Argument-3> attacks a premise of <Argument-1> that supports ”Statement” (i.e. that concludes that s). <Argument-4> attacks a premise of <Argument-2> which rejects ”Statement” (i.e. claims that non-s). The arguments 3 and 4 do not have a direct dialec-

[2]Additionally, the graphical representation of the dialectical situation from Figure 1 is slightly modified: a premise from <Argument-2> (A-2-Premise-1) is omitted from the graph for reasons of presentation.

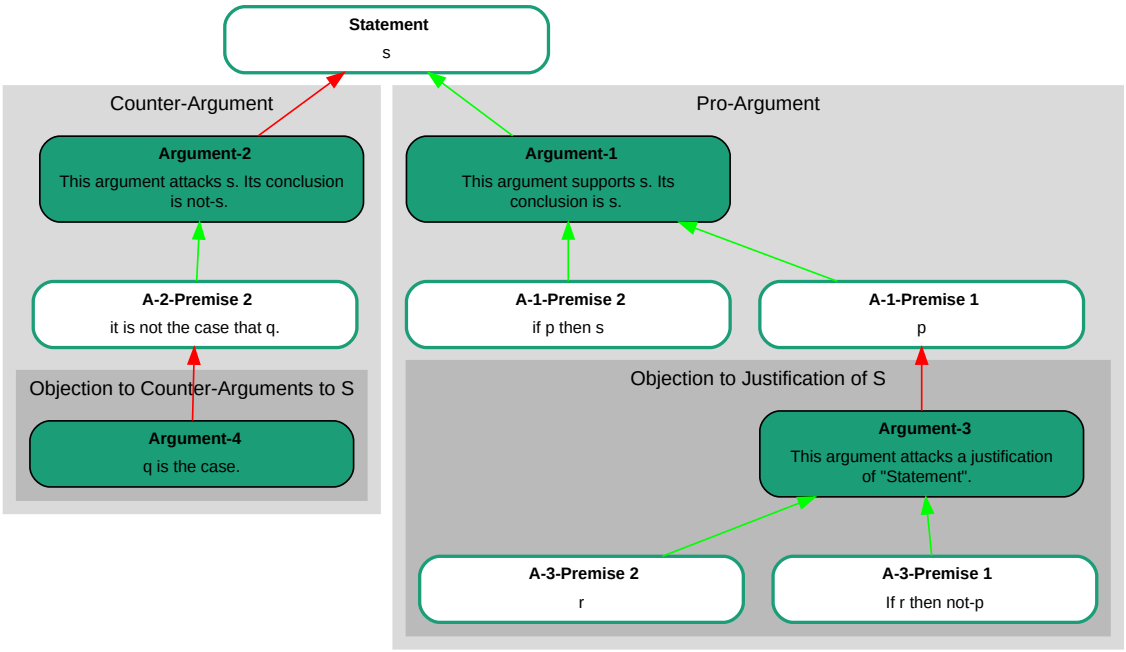

Figure 2: Example of a Dialectical Structure

tical relationship to the statement of interest that s: <Argument-4> attacks an objection to the claim that s. If <Argument-4> holds (i.e. all its premises are true), then the objection to ”Statement” (contained in <Argument-2>) does not hold. This, in turn, does not justify the statement that s, it merely shows that an objection to it does not hold. Similarly, <Argument-3> attacks a justification of the statement that s. If <Argument-3> holds, ”Statement” lacks a justification in this dialectical situation. But that does not mean that the statement that s is wrong (or implausible or not acceptable).

Evaluation of Arguments Reconstruction of the pro- and counter-reasons in terms of deductively valid arguments – as described in the previous section – is the first step of an argument analysis. The second step is to evaluate the epistemic value of the reconstructed premises. Here, the task is to analyse what epistemic value is transferred from the premises to the conclusion: if all the premises are true then the conclusion must be true as well; if at least one premise has a weaker epistemic value above being false (e.g. ”plausible”), the conclusion has maximally this weaker epistemic value; if at least one premise is false, the conclusion is not justified (no epistemic value has been transmitted, so to say).

The task of assessing the epistemic value of premises may require empirical expertise (in case of evaluation of premises with empirical content) or an appropriate normative reflection in case of a premise with contested normative content. In this report, we will not provide a complete evaluation of all reconstructed premises. Instead, we will point out the premises which play a dialectically relevant role and whose epistemic value is obviously wrong or very weak or whose epistemic value is undecidable

for us.

With the argument analysis, we do not pursue to evaluate all arguments directed at the use of TWONs. Instead, our goal is to identify loopholes in the justifications for the use of TWONs and for its limitations. Based on these outcomes, we intend to suggest ways in which research might proceed in the light of these loopholes. So, let us turn to describing how we use the argument analysis for an assessment of TWONs.

2.2 Application to the case study: justification of deployment of TWONs

As we have indicated in Section 1, it is reasonable to expect undoubtedly desirable outcomes from the availability of TWONs – if their development is successful, they will be a game-changing tool for taming OSNs – and, at the same time, it is reasonable to expect grave societal risks from their deployment. These forecasts raise the question of whether a TWON should be developed and deployed. The following answers to this question are possible:

- [T1]: TWON should be researched and developed.
- [non-T1]: It is not the case that TWON should be researched and developed. An equivalent formulation of non-T1 is: Research and development of TWON should be prohibited/prevented.
- [T2]: TWON should be publicly available.
- [non-T2]: It is not the case that TWON should be publicly available. An equivalent formulation: Public access to a TWON should be prohibited.

How can the four claims T1, Non-T1, T2, and Non-T2 be justified in the light of the expected benefits and risks of a TWON? Intuitively, one would expect that a certain kind of ”balancing” between the expected benefits and the foreseen risks is required. However, without a deeper dive into this issue, it remains unclear what this ”balancing” consists of and what its rationale might be. Here, an argument analysis can be informative. Using argument analysis, we will identify plausible arguments which (deductively) bridge the forecasts about TWONs potentials and risks with the practical recommendation regarding its use and development. Figure 3 depicts this task in form of an argument map.

The arguments we are looking for – depicted in Figure 3 in the argument-boxes in the green group ”Arguments to be identified by the argument analysis” – contain potential risks and benefits from TWONs in their premises and one of the main statements of interest – T1, non-T1, T2, non-T2 – as their conclusions. The premises which need to be added to the arguments <A1> to <A4> will, therefore, contain ethical evaluations or principles which allow a logically valid derivation of the statements of interest. Evaluation of their plausibility will demonstrate how plausible justifications for the statements of interest are.

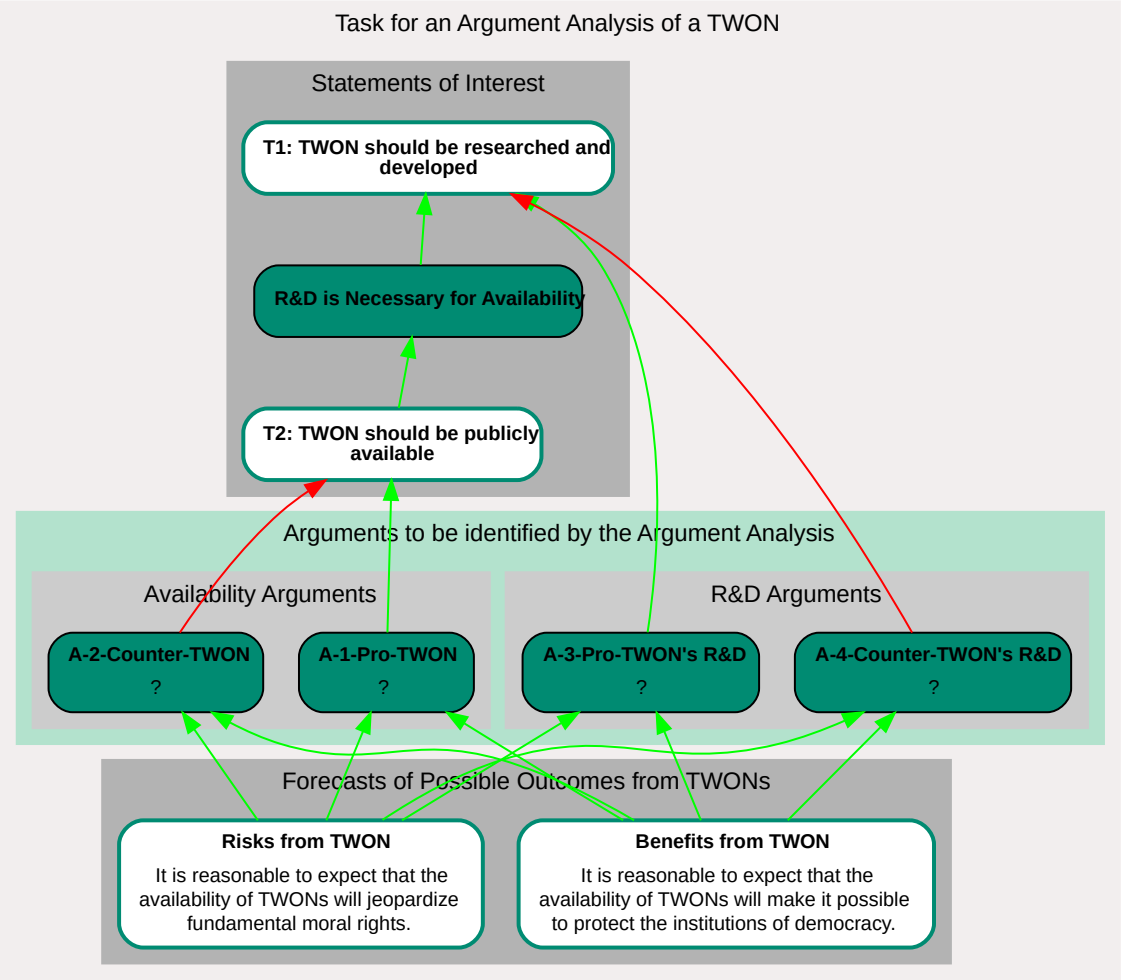

Figure 3: The Task of the Argument Analysis

The gray box at the bottom contains the statements that TWONs are expected to bring about important benefits as well as high societal risks. The gray box at the top contains the normative statements of interest T1 and T2 (their negations are not depicted). Filling out the boxes within the green box in the middle is the task of the argument analysis. Currently, they contain dummies for arguments which support and attack the statements of interest. We depict that exactly one argument supports each of the main statements. This is an assumption for the sake of representation – there might be several arguments supporting the main claims.

Note that there is a logical relation between the claims T1 and T2: Somebody who accepts T2 (i.e. believes that TWON should be publicly used) must also accept T1, i.e. the demand for its research and development just because research and development are necessary for availability of a TWON. This relationship is represented by the connection of T2 and T1 via a green arrow which means that T1 is entailed by T2. The argument-box <R&D is Necessary for Availability> represents the argument with T2 as one of its premises and T1 as its conclusion.

There is a huge number of possibilities how a deductively valid argument can connect the statements about TWONs' risks and benefits to the statements of interest. We do not aim at identification of all meaningful arguments. Rather, our goal is an adequate construction of the most relevant arguments in the sense that they could be brought forward by some stakeholders. The endeavour of an adequate construction of practically relevant arguments involves "creative and normative moves" (Brun and Betz, 2016, p.43). The creative move in argument construction consists in imagining ways of reasoning for a certain claim. The normative move consists in the choice among premises. Let us visualize this point with an example which will occur in the constructed arguments in Section 6.

Consider the following reasoning: 'TWON has high potential for protection of democratic norms. Therefore, it should be developed.' The reason giving idea can be interpreted in two ways, as a deterministic forecast or as an statement about one's expectations:

(1) [Deterministic Forecast]: The availability of a TWON will enable a democratic society to protect its democratic norms.

(1*) [Articulation of Expectations]: It is reasonable to expect that the availability of a TWON will enable a democratic society to protect its democratic norms.

The choice of the formulation of the first premise influences further premises of a deductively valid argument. Consider the two resulting arguments:

Argument-1: Deterministic Forecast

(1) The availability of a TWON will enable a democratic society to protect its democratic norms.

(2) If a technology T enables a society to protect its democratic values, T ought to be developed if there are no other means for protection of democratic values and T does not involve additional moral drawbacks.

(3) There are no other means for protection of democratic values besides developing a TWON.

(4) The availability of a TWON will not involve additional moral drawbacks.

---

(5) A TWON ought to be developed.

Argument-2: Articulation of Expectations

(1*) It is reasonable to expect that the availability of a TWON will enable a democratic society to protect its democratic norms.

(2*) If it is reasonable to expect that a technology T enables a society to protect its democratic values, T ought to be developed if no other means for protection of democratic values are known at t and there are at t no reasons for the expectation that T will involve additional moral drawbacks.

(3*) No other ways to protect democratic values besides developing a TWON are currently known.

(4*) Currently, there are no reasons for the expectation that availability of a TWON will involve additional moral drawbacks.

(5*) A TWON ought to be developed.

Premise (1*) is much easier to justify than (1). The latter is hardly justifiable: as fallible beings, we cannot deterministically forecast far future events. However, the principle in (2) that allows a conclusion to (5) is true: it is not possible that all the conditions in the antecedence of (2) are met and that it would be morally wrong to develop T.[3]

However, the formal principle in (2*), which is needed because of the epistemic formulation of the first premise (1*), is, strictly speaking, false: it is possible that there are good reasons to expect that T will lead to desired consequences and that the other conditions of (1*) are met, but that it is morally wrong to develop T because T, contrary to our well justified expectations, will lead to devastating consequences.

Both reconstructions of the reasoning contain drawbacks: <Argument-1> presupposes an empirical premise (1) which is hardly justifiable but a formal principle which is true. <Argument-2> presupposes a plausible empirical premise (1*) but must presuppose a formal principle which is, strictly speaking, false. To avoid reconstructing a large number of arguments with different drawbacks for one and the same reasoning, it is necessary to choose the most appropriate reconstruction. Different interpreters can consider different reconstructions as appropriate.

Often, the results of an reconstruction of a reasoning are not unique: usually, there are several options for an adequate reconstruction of a reason formulated in a common language. However, a reconstruction is not arbitrary, neither. Our goal is identification of arguments which belong to the set of adequate reconstructions of relevant arguments. When we make choices, we will make them transparent and justify them.

2.3 Steps of an Argument Analysis

There are several conceptual, empirical, and normative clarifications necessary before an argument analysis can begin.

Step 1: Stage Setting The guiding question with which we have started asks which controversial ethical assumptions are involved in the process of development of a TWON. So far, we have proposed to identify these assumptions by reconstructing arguments justifying a proposition that TWON's development and use are allowed and a proposition that it is not.

However, there are different options how a technology can be "used" in a society. This depends on how a society regulates access to a technology. The practical statement of interest T2 in Section 2.2 – `TWON should be publicly available` – represents only one option for how a society could

[3] The proposition in (2) has the logical form of a conditional: $p \rightarrow q$. It is false if $p \wedge \neg q$. Since it is not possible that $p \wedge \neg q$, (2) must be true.

regulate its access. In this case it is an extreme: free availability of a TWON. Another extreme option is a way of governance of TWONs according to which a very few members of a society obtain access to the technology. In between is a continuum of possible policy options all of which influence the resulting benefits and risks from TWON. Thus, the first step is to clarify the relevant modes of regulation of a TWON and the resulting practical statements.

Outcome Step 1: List of the relevant modes of regulation of a TWON with a specification of the practical statements of interest.

Step 2: Collection of Empirical Data Evaluation of a technology requires knowledge about how this technology interferes with the empirical world. Usually, this interference can be structured in terms of hte required inputs and the expected outputs:

- Required Inputs: Estimation of all relevant inputs necessary for research, development, and maintenance of a TWON for all governance modes. Examples for relevant inputs are: financial cost, natural resources, data.
- Expected Outcomes: Identification of all reasonably expected and relevant consequences from all applications of a TWON for all governance modes (based on outcomes in Step-1).

Outcome Step 2: List of the required inputs and the expected outputs for all modes of regulation.

Limitation: An empirical assessment of the required inputs and the expected outputs of a TWON lies beyond the scope of this report. Instead, we will collect categories of inputs and outputs mentioned in the literature (cited in Section 4) without quantifying their size. Based on this qualitative description of inputs and outputs, we will identify controversial ethical claims. Their controversies might result from empirical uncertainties. This outcome will then pave a way for the next round of assessment with a more precise focus on which additional data should be obtained.

Step 3: Prima Facie Reasons Since the task of this analysis is to construct arguments that justify general moral statements about a technology (”Technology T ought or ought not to be used”), we need to clarify the moral significance of the technology’s inputs and the consequences from its deployment. If they are morally significant, they provide prima facie reasons in favour or against a development or use of the technology. Prima facie reasons can be understood as pieces of a jigsaw puzzle. A prima facie reason in favour of p is not decisive, i.e. it does not imply that p should be done or chosen. It could be outweighed by another reason or it could be silenced in the light of another reason[4]. All prima facie

[4]Philosophers distinguish between prima facie and pro tanto reasons which play a slightly different role in determining a practical conclusion (Scanlon, 1998, p.50f.). These differences are irrelevant here. We use the term ”prima facie reasons” for convenience referring to pro tanto reasons as well.

reasons directed at an action determine whether the action ought to be undertaken or not. In this step, all the pieces of the puzzle – the pros and cons – are collected.

Outcome Step 3: List of the pro- and counter-reasons for all modes of regulation.

Step 4: Controversial Ethical Commitments Based on the results from Step 3 – List of the pro- and counter-reasons for each mode of regulation – the argument analysis will be conducted.

Outcome Step 4: Controversial ethical claims for the justification of all governance modes

Summary 2

The aim of this handbook is to inform a debate about and a decision on whether and how a TWON should be developed and deployed. We point to open questions that research and ethical deliberation needs to address before an informed decision about TWONs can be made. For that, we identify core arguments pro and con TWONs and the premises that these arguments hinge on.

What to **expect** from this report:

- controversial arguments justifying or rejecting alternative modes of governance of a TWON
- recommendations for how to deal with ethical controversies in the research process
- methodology for identification of ethical controversies in the further course of research

What **not to expect** from the ethical analysis in this report:

- Recommendation for one of the modes of governance of a TWON
- Completeness with regard to reconstructed arguments
- Empirical research

3 Step 1: Stage Setting

A TWON is a computational model which reliably replicates dynamic behaviour of an online social network (OSN). A TWON models users of OSNs as nodes in a network where network links connect users who may communicate with each other. Next, there is a so-called ”user model”, to describe how users behave on the network. The user specifies all user decisions about becoming active, such as selecting

content, producing content, as well as evaluating and sharing content. Finally, a TWON has a ”platform model” describing the structure users are interacting on. This part of the TWON contains all algorithms and platform affordances guiding what content users are exposed to and when. For instance, personalization algorithms ranking incoming content are part of the platform model.

Here, we make the following assumptions about TWONs:

- The TWON-model reliably represents the structural aspects of the real network of the target OSN
- The TWON-model reliably represents the communicative structure of the target OSN (communicative behaviour of the OSN-participants according to which they decide to contribute content on an OSN and what to contribute)
- The TWON-model reliably represents the platform algorithms of the target OSN.
- More generally, we assume that the TWON reliably replicates content of the target OSN. We can think of that as a model, which, after being trained with the data from an OSN from a time period t_1, reliably replicates the content which has been produced on the target OSN in a time period t_2.

These assumptions are certainly heroic in the sense that there is a lot of work to be done until they are met. Moreover, these assumptions hide ethical and epistemic challenges which arise before the reliable TWONs will be available. Validation of agent-based-models (ABMs) which aim to represent social systems is notoriously controversial (c.f. Klein et al., 2018, p.16ff.). Moreover, these ABMs will be programmed with decision-making algorithms and they will create content by language generating algorithms. These algorithms bring their own epistemic challenges (Mittelstadt et al., 2016): conclusions of algorithms depend on the quality of the data from which they are derived. The data can be inappropriately biased or insufficient. For machine learning algorithms, these challenges aggravate since it remains opaque by which inferences they derive their conclusions from the background data. Thus, an important task in the development of TWONs will be to propose epistemic criteria for assessing the validity of model results. The lack of an assessment of the validity of models simulating online social networks will lead to ethical challenges as soon as the models are used for public policy purposes: as long as an epistemic interpretation of their results remains unclear, they can be used to justify morally problematic policy choices.

By focusing on an ideal TWON, we do not mean to downplay the epistemic and ethical obstacles that will arise much earlier. Rather, the goal of the present analysis is to identify ethically controversial concerns on the way of development of a certain technology. For that reason, we start with the ideal outcome of technological research we are interested in and analyze the question of which objections can be articulated against the ideal outcome of the research process.

For our analysis, we do not presuppose a specific OSN. It could be any large online social network from the European Commission's list. However, we distinguish four scenarios for how a TWON could be governed, we call them "governance modes" (GMs). The outcomes from the use of TWONs differ depending on how a society organizes their governance, i.e. how it regulates who owns the model (i.e. maintains and regulates the access to it), who has access to it and its conditions. The GMs we distinguish span a continuum of possible governance modes of a TWON.

GM-1: unrestricted public access to the TWON This GM represents the maximally free public access to the TWON.

- Ownership: a publicly financed institution runs and maintains the model: it updates the code and the background data. We assume that the code of the model is publicly published under an open source license. The TWON governing institution has the task to ensure that the TWON is used in compliance with the current laws.
- Access: public users can access the model. They can insert their queries for a model run and get a model output.

GM-2: Public Interest Authority governs the TWON This GM represents the idea that a publicly authorised institution regulates the access to the TWON to ensure that the TWON is used only in the public interest.

- Ownership: a publicly financed institution – the governing authority – runs and maintains the model and ensures that only the authorized individuals access the model. Only the governing authority has access to the code and the background data.
- Access to the TWON is only permitted to investigate matters of public interest. A legislative body must define what constitutes the public interest. The result could be a list of topics for which a TWON can be used, as, for instance, the Digital Services Act of the EC specifies a list of systemic risks (DSA, Chapter 4, §34(1)).

  We assume that the governing authority regulates the access to the TWON in the following way: all organizations that pursue an objective that is in line with what has been defined as the public interest will be granted access to the TWON. The organizations decide how they distribute the access rights internally (whether they provide access only to certain members or to everybody). An organization can lose its access right in case its representatives misuse it. Depending on how the "public interest" is specified, a more or less broad set of organizations will have access to

the TWON. At least the following institutions will have it: political organizations (executive, legislative, juridical bodies, parties, non-governmental organizations etc.), academic institutions, certain non-profit companies and clubs of benefit to the public (in German: ”gemeinnütziges Unternehmen”) whose purposes comply with the specified public interest.

GM-3: Only Vetted Researchers have access to the TWON This GM represents the idea that access to the TWON is limited to researchers only if their research contributes to the specified goals. The TWON is governed by a research institution mandated for this task (analogously to other sensible research instruments: data panels (e.g. the socio-economic panel SOEP), CERN etc).

- Ownership: a research institution which hosts, maintains, and updates the TWON. Only the hosting institution has access to the code and the background data. It regulates who is allowed to use the TWON.
- Access to the TWON is restricted to researchers for enquiries in the public interest. The public interest is defined by a legislative body.

  Researchers apply to the hosting institution to use the TWON for a specific research query. The proposal explains in how far the intended research contributes to the public interest and takes a stand on ethical concerns as, e.g., sensible data requirements. The hosting institution decides after evaluation of the research proposal. If a research question requires sensible personal data, the decision committee must weigh between the relevance of research and the protection of personal data.

GM-4: Access to TWONs on behalf of a Judicial Authority This GM represents the mostly restricted way of governance of a TWON. Access to it is regulated by a legal authority. In case of a serious breach of the law in which communication on an OSN might have played a causal role, a court may commission a group of investigators to analyze by a TWON of the OSN whether the respective communication on the OSN has contributed to the law break.

- Ownership: a publicly financed institution which hosts maintains, and updates the TWON. Only the hosting institution has access to the code and the background data.
- The TWON may only be used by order of a judicial authority. The legal authorities decide in accordance with the law in which cases they order enquiries by a TWON.

Figure 4 summarizes the governance modes assumed for this analysis.

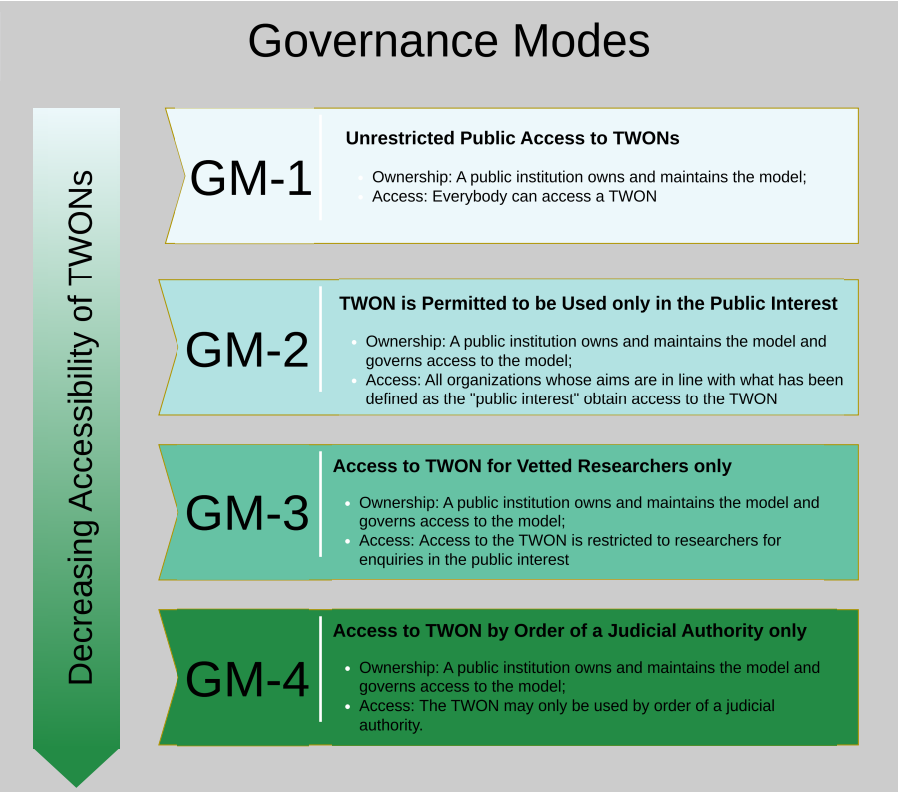

Figure 4: Overview of the Governance Modes

4 Step 2: Identification of the Required Inputs and Possible Outputs

The technology assessed in this report – a TWON – is currently at an early stage of its development. At this stage, prediction of outcomes from the availability of TWONs is highly uncertain. Outcomes can currently be predicted in general categories without quantifying their magnitude. Similarly, the inputs can be structured into types of required inputs without their amount being reasonably quantifiable. In this step, we firstly systematically review the categories of the inputs needed for TWONs research and use and the categories of the expected outputs. The results are based on developers' estimations and a literature review (Sobkowicz, 2017; Gilbert et al., 2018; Margetts and Dorobantu, 2023; Mihai et al., 2022). Additionally, we show the differences in inputs and outputs between the scenarios (c.f. Section 3) based on the current state of knowledge.

4.1 Required Inputs

A TWON will require four types of resources for its functioning:

- material resources from which the computational hardware is built and electricity for hardware's operation is generated;
- human labour for developing, producing, and operating the TWON;
- data from the target OSN which the TWON mimics: data about agents which communicate on the OSN and data on the content they communicate (read and produce);
- financial resources for acquisition of material resources and, if applicable, of data.

Recent assessments of resource use in the field of artificial intelligence (Bommasani et al., 2021; Strubell et al., 2019; Wu et al., 2022) suggest that the development and deployment of applications in this field require significant physical, human, and financial resources. Since TWONs will deploy AI-tools such as natural language processing models, its development and deployment might become resource-intensive. At the current stage of development, most of the decisions that determine the resource use are not made. Therefore, the resulting ranges of possible input requirements would be very wide. Environmental impact analysis would be reasonable on the basis of first blueprints for TWON prototypes.

Regarding the data requirements, TWONs face a challenge which distinguish them from other AI applications. As a TWON replicates a target online social network, it needs the original data from the respective network to function reliably. As we described in the introductory section (1), data requirements raise a trade-off: the more personal data from the target OSN is available as input for the TWON, the better are the outcomes of the TWON.

The amount of the required data cannot be specified at the moment. But we distinguish three scenarios for how the data could be acquired:

- **Scenario 1 "Full Data Set"**: the TWON has been trained on a sufficiently large set of complete data from the target OSN. The data set has been provided by the OSN-owner on the legal basis of the European Commissions Digital Services Act providing researchers access to the data of the very large online platforms for specified research purposes (c.f. Article 40 and 34 of the DSA).
- **Scenario 2 "Collected Data"**: the TWON has been trained on a set of data from the target OSN freely available on the internet and collected by the developers of the digital twin. We assume that the collected data set is sufficiently large to train the machine learning algorithms of the TWON such that the model fulfills the criteria of a digital twin (it allows reliable prediction and analysis of outcomes on the target OSN at a macro level).
- **Scenario 3 "Informed Consent Data"**: collection of data from a set of individuals who have agreed to provide the data on their communicative behavior on the target OSN for the purpose

of developing and using a model of the OSN (the individuals agree to act as a kind of “sensors”) after having been sufficiently informed about the purposes of the TWON. Again, we assume that the collected data set is sufficiently large to train the machine learning algorithms of the TWON.

4.2 Possible Outputs

Our estimation of possible outcomes from the deployment of TWONs rests on the assumption that a TWON successfully replicates its target OSN and provides a reliable virtual copy of the OSN which can be used for testing of various parameters.

- (1-1) **Detection** of communicative outcomes on the target OSN which are of particular interest (e.g., undesired outcomes such as hate speech)
- (1-2) **Attribution**: counterfactual analysis (simulation runs with systematically varying assumptions) allows identification of factors which have causally influenced an outcome of interest on an online social network. If communication on an OSN has caused policy changes, the TWON allows attribution of causally contributing factors to the policy change as well.
- (1-3) **Measurement**: a virtual simulation of an OSN allows scientific analysis (measurement: structured representation and quantification) of the respective online social network (Lazer et al., 2021; Margetts and Dorobantu, 2023).
- (1-4) **Prediction**: forecasts of communication dynamics on a target OSN.
- (1-5) **Manipulation** Analysis of causally contributing factors (Outcome (1-2)) and hypothetical prediction (testing of outcomes of systematic interventions on the target OSN) allows identification of instruments (communication strategies) by which opinion dynamics on the target OSN is changed in an intended direction.

The scenarios for how a TWON can be governed in a society (Section 3) have an impact on the outputs and the amount of the required inputs. Table 1 depicts the differences between the scenarios.

Types of possible use of a TWON	**GM-1**	**GM-2**	**GM-3**	**GM-4**
(1-1) **Detection** of communicative outcomes on an OSN	any kind if of interest for a societal agent	outcomes of interest for organizations pursuing public goals	only outcomes considered as relevant by the TWON governing institution	only illegal outcomes
(1-2) **Attribution** causal factors of communicative outcomes				only causal factors of illegal outcomes
(1-3) **Measurement**: new knowledge about social systems		research in any field of interest will be conducted	only research which complies with the specified goals (topics of public interest)	no research with TWON expected
(1-4): **Prediction** of communication dynamics on an OSN		future outcomes of interest for organizations pursuing public goals only	only outcomes which comply with the specified goals (public interest)	no prediction expected
(1-5): **Manipulation** Identification of instruments for influence of opinion dynamics on an OSN in an intended direction.		public interest organizations will identify means for directed influence of opinions	only means for intended influence which comply with the specified topics of public interest	no outcomes expected

Table 1: Consequences from deployment of TWONs according to the governance modes

According to the assumptions of the governance mode 1, there are no restrictions on TWON access. Thus, any societal agent with the necessary epistemic or financial capacities can use it for its own interests. All societal agents engaged with public communication – commercial enterprises, public policy stakeholders, academics – have a reason to use a TWON to improve their outreach: by detecting communicative outcomes of relevance for their publicity, for analysing their causal factors, by developing instruments for intended influence of opinions related to their commercial or political interests. Since a TWON can support a wide variety of societal agents in pursuing their respective interests and since we assume in this governance mode that the epistemic and financial costs of its use will be rather low, it is reasonable to expect a very broad deployment of the technology by a broad variety of societal members.

In scenario GM-2, access to the TWON is restricted to members of organizations pursuing public goals. Accordingly, the number of societal agents with access to the TWON is smaller than in GM-1. Particularly, stakeholders pursuing commercial interests only will have no legal access to the TWON. Still, all societal agents pursuing public goals – public policy stakeholders (politicians, governmental and non-governmental organizations), academics, non-commercial enterprises, clubs etc. – will have access to it. All of them who are engaged in public communication will have a reason to improve their outreach using a TWON. Academics will have access to the TWON for research activities. Thus, it seems to be reasonable to expect that a wide majority of stakeholders pursuing public interests will use the TWON.

Scenario GM-3 restricts access to the TWON to researchers with a specific project which needs to be approved by the TWON governing institution.Deployment of the TWON for detecting communicative outcomes or of their causal factors is restricted to projects that aim to advance knowledge in one of the fields which have been declared as relevant by a legislative body. Similarly, it is conceivable that some research projects will identify means for manipulation of opinions on an OSN if the governing institution decides that this complies with the declared fields of research. The main use of TWONs in this scenario will be to measure social networks (1-3) with the aim to advance knowledge in the field of interest.

Finally, according to the scenario GM-4, the TWON will only be used to detect illegal acts on OSNs or in the real world, if the latter have been caused by communicative acts on an OSN and to clarify causal factors that have influenced an illegal outcome. All other possible applications of a TWON are not legal according to this governance mode.

5 Step 3: Evaluation of Inputs and Outputs

This step aims to identify prima facie moral reasons for and against deployment of a TWON according to a particular mode of governance. This task involves an evaluation of the TWON's inputs and outputs in terms of their moral significance. Such an evaluation presupposes an ethical background theory (which justifies the evaluative conclusions).

Philosophical literature does not offer a standard ethical background theory to use in practical evaluations. Instead, there are several grand ethical theories (consequentialist, deontological, and virtue ethics theories) providing different accounts of the sources of normativity and, correspondingly, differing in resulting axiologies. In certain, often very specific cases, different grand ethical theories recommend opposing practical conclusions (i.e. statements about whether an act is morally permissible or not)[5]. However, there is also a broad range of agreement between the theories. Since we are discussing an evaluation from a highly general perspective with the aim to identify ethical and empirical issues which deserve a more detailed analysis, the differences between the background theories should not matter to our assessment.

Therefore, we do not presuppose a specific ethical background theory. Rather, we develop a taxonomy using what might be termed a pragmatist approach. Our starting point is the moral problem which motivates the research endeavour (c.f. Section 1). Spread of mis- and disinformation, emergence of bubbles and echo chambers, foreign inferences in election processes provide reasons for concern that democratic governance (Persily and Tucker, 2020) and individual autonomy (Susser et al., 2019) are undermined. Individual autonomy is an ability to make self-determined decisions. We can think of them as decisions which an individual would make if he/she had complete background information on the topic and enough time for contemplation on the pro- and counter-reasons, taking into account various evaluative perspectives (Susser et al., 2019, p.8).

It is controversial in ethical background theories whether the democratic constitution of societies constitutes an own moral value or is morally valuable because it is constitutive of other basic moral values – for example, individual autonomy or well-being. Since it is a reasonable position that the values "democratic governance", "individual autonomy" and "well-being" are distinct[6], and we would prefer to make the mistake of double-counting moral values instead of the mistake of missing their moral significance, we treat them as distinct and include them all into our value matrix.

An additional moral concern motivating this analysis is the need to use personal data to develop

[5] Cases of the so-called "double effect" (causing harmful side-effects by promoting good) provide examples for where the grand theories justify contradictory practical conclusions (c.f. McIntyre, 2023).

[6] It is not unreasonable to accept trade-offs between these values: One might prefer a social world with lower well-being and higher democracy ranking to a social world with higher well-being and lower democracy ranking.

and deploy the TWON. The more personal data the TWON is based on, the more realistic will be its outcome (c.f. Section 1). The use of personal data collected through digital services raises moral concerns because it may violate what has been called the ”right to privacy” or the ”right to informational self-determination” (Moore, 2003; Roessler, 2017). The latter refers to individuals' ability to control ”access to information about themselves” (van den Hoven et al., 2020). Again, there is no agreement on whether the right to privacy constitutes an own moral right or is an instrument for other, more fundamental moral rights (such as autonomy, democracy, freedom, well-being etc.).

The ethical topics motivating this research can be structured in the following way:

(1) Welfarist Values:

  (1-1) social welfare

  (1-2) individual quality of life

(2) Non-Welfarist moral values or norms which are known to be undermined by the spread of communication on OSNs:

  (2-1) the right to privacy or informational self-determination: the ability to control which personal information is accessible to others;

  (2-2) individual autonomy: ability to make self-determined decisions

  (2-3) ideals of deliberative democracy: political power in a society is legitimized by individual votes formed in a process of deliberation among citizens.

The moral categories distinguished so far do not cover all moral concerns that might be affected by a technology. For instance, technologies have often effects on the distribution of epistemic abilities, financial resources, social recognition of different groups etc. To accommodate for these effects as well, we include an additional category into our taxonomy matrix:

(3) All other moral values and norms

Again, the above limitation applies: the assignment of particular values and norms to the moral categories differs with regard to the underlying ethical theory and a more precise specification of the values. For example, the distribution of epistemic abilities or the recognition of social groups can be considered as necessary conditions for the ideal realization of deliberative democracy. Therefore, one might argue, these values should be assigned to the second category and not to the third. We suggest to postpone those debates to a later point (concerning the mentioned example: when it comes to a more precise specification of the impact of TWONs and OSNs on the ideals of deliberative democracy). The

Pro-Reasons GM-1	Counter-Reasons GM-1
(B-1-1) Increase in social welfare (in an amount of ΔSW(GM-1))	
(B-1-2) Improvement of individual quality of life (in an amount of ΔQoL(GM-1))	(R-1-2) Risk to individual quality of life from misuse of personal data
	(R-2-1) Increased violations of the right to informational privacy
(B-2-2) Ability to prevent undermining of individual autonomy on OSNs	(R-2-2) Intensified undermining of individual autonomy
(B-2-3) Improvement of the deliberative quality of public discourse, but at least the ability to prevent the erosion of epistemic, moral, and juridical norms necessary for democratic governance	(R-2-3) Risk of collapse of institutions necessary for democratic governance
(B-3) Additional knowledge for mitigation of existing injustices (e.g., marginalization of vulnerable groups)	(R-3-1) Harms to future generations from increased resources use and/or environmental pollution
	(R-3-2) Reinforcement of existing inequalities in financial and political power relations at the national and global level

Table 2: Prima facie reasons regarding an unrestricted access to the TWON

purpose of the axiology presented here is to guide us in structuring pro- and counter-reasons without omitting one of them. The suggested axiology should be sufficient for that purpose.

Since we do not presuppose any particular ethical background theory, we will be flexible in our use of ethical terms. We will use the terms ”benefits”, ”advantages”, ”pro-” or ”supporting reasons” interchangeably; similarly, we will use the notions ”risks”, ”disadvantages”, ”counter-reasons” equivalently.

5.1 Evaluation of the GM-1: Unrestricted Access

Table 2 provides an overview of the morally significant advantages and drawbacks from an unrestricted access to a TWON.

5.1.1 Pro-Reasons (Benefits) GM-1

Welfarist Values As described in Section 4.2, agents engaged with public communication (political, commercial, academic stakeholders) have an incentive to use TWON for developing better strategies for

their outreach. It is reasonable to expect that an unrestricted access to the TWON will thereby increase economic product on aggregate. For commercial agents will be better able to align their products and services to individual's preferences. Currently, knowledge from OSNs on individual preferences is monopolized by the OSN providers. With a free access to TWONs of OSNs, this knowledge will become freely available thereby destructing the deadweight loss from monopoly.

Free access to the TWON allows agents with different interests to experiment with complex social systems and their dynamics (O-3: Measurement). This amounts to a groundbreaking advance in the study of complex social systems from which it is not unreasonable to expect fundamental advances in knowledge. These advances would, in turn, be used by commercial or public organizations to develop new products and services that improve various aspects of individual quality of life and generate economic benefits for their developers ((B-1-1) and (B-1-2) in Table 2).

Non-Welfarist Moral Norms Jeopardized by OSNs Stakeholders interested in protecting individual autonomy obtain an opportunity to identify design decisions in an OSN that threaten self-determined judgement and to develop measures against undermining individual autonomy ((B-2-2) in Table 2).

Stakeholders interested in protecting norms and values necessary for democratic governance that are threatened by OSNs will be able to use the TWON to predict problematic communicative outcomes (hateful and abusive forms of communication, mis- and disinformation, attempts to intentionally influence public opinion), identify their origin, and to develop instruments for their prevention. Currently, private companies owning OSN platforms determine which norm violations they regulate. With a free access to TWONs, all potentially affected stakeholders could identify norm violations in advance and determine causal effects of observed misbehavior. Moreover, free access to TWONs enables interested stakeholders ito develop tools to improve the deliberative quality of online communication – be it design of content distribution algorithms, be it additional applications supporting access of information for a sufficiently informed and a well-reflected judgment. In the best case, these initiatives will lead to an overall improvement of the deliberative quality in a society (B-2-3 in Table 2).

Other Moral Values and Norms Since a free access to a TWON might lead to groundbreaking advances in understanding of complex social systems, it is reasonable to expect that this technology will be used to address various societal challenges in addition to threats to democracy. For instance, knowledge could be provided on the effects that lead to the marginalization or status changes of social groups. This, in turn, would enable development of strategies for mitigation of these injustices ((B-3) in Table 2).

5.1.2 Counter-Reasons (Risks) GM-1

Risks for Welfarist Values We cannot reasonably justify the claim that a free access to the TWON imposes risks for societal welfare (in terms of an aggregated economic product). However, this governance mode carries risks for individual quality of life. Since a TWON replicates an OSN and an OSN contains various personal data, it is principally possible to use the TWON to elicit personal data which then can be used in a way harming an individual. Such a use of TWONs will be illegal. Still, due to a free access to the technology, it would be challenging to enforce this legal norm ((R-1-2) in Table 2).

Non-Welfarist Moral Norms Jeopardized by OSNs A functioning TWON will be based on personal data from participants of a target OSN. Whether this creates a prima facie reason against TWON depends on two criteria: (i) whether the data have been acquired in a morally legitimate way (fair acquisition of a resource); (ii) whether the use of personal data for development and/or operating of the TWON violates the right of individuals to informational self-determination. Currently, it is unclear what gives an individual the ability to control personal data (Mittelstadt et al., 2016; van den Hoven et al., 2020). For this reason, it remains undecidable in many cases whether the use of personal data violates moral norms.

In Section 4.1, we have suggested three scenarios for how personal data could be acquired. The third scenario – data are collected from individuals with their consent after being sufficiently informed about the purposes of the data use – does fulfill the two criteria for a morally legitimate data use. But it is also a scenario which is practically difficult to meet. For the other two scenarios – (i) data provided by the OSN owner on behalf of a legal norm and (ii) freely available data collected – it remains undecided if their use is morally objectable. In both scenarios, one might argue that the acquired data violates the individuals' right to informational self-determination. The individuals are not aware that the data they left via using certain internet applications will be used to develop a technology with various commercial applications and certain risks, theoretically even for themselves[7]. Against this, it might be objected that it is unrealistic to require individuals to be informed about all potential uses of their data collected through their use of digitally connected devices (Froomkin, 2019). Because of this disagreement, we evaluate the counter-reason from violation of the right to informational privacy as a possibility: it might turn out to be a justified worry ((R-2-1) in Table 2).

It is reasonable to expect that an unrestricted access to a TWON will increase communicative activities which undermine individual autonomy and threaten democratic governance ((R-2-2) and (R-2-3) in Table 2).

[7] There is evidence that individuals are not aware of possible uses of their digital data and disagree with its (non-consented) use in research (c.f. Fiesler and Proferes, 2018).

Some societal agents are interested in public's beliefs on a certain topic deviating from the best justified beliefs on that topic. For example, some commercial agents are interested in consumers believing that their products are the best ones independently of whether this is true or not; some political agents are interested in voters believing that their political position ought to be supported at the ballot; they are not interested in voters finding out which political position is best justified. The TWON will offer these agents a more effective tool for a more precisely targeted spread of mis- and disinformation thereby undermining individual's ability to form justified beliefs.

A more precisely targeted spread of mis- and disinformation and more effectively tailored attacks that seek to distort public opinion from the best justified position can, in the worst case, lead to the collapse of fundamental institutions necessary for democratic governance. Since the TWON would drastically increase the precision of manipulation of public opinion and its public availability will enable a wide access to it, we consider this thread as one which should be reasonably expected.

Other Moral Values and Norms Unrestricted access to TWONs might lead to a very high demand for the use of this technology. Its use is relatively resource intensive. In the worst case, widespread use of TWONs could lead to an overall increase in scarcity of critical non-renewable resources needed to produce the hardware, or in emissions of greenhouse gases to provide the energy for the runs of TWONs – both consequences harming future generations ((R-3-1) in Table 2).

It is not unreasonable to expect that an unrestricted access to a TWON will reinforce existing injustices. Politically and economically most powerful societal agents will use the TWON to influence public opinions in a way that their political and/or economic position is sustained and hinder other societal agents from its deployment. Change in political and economic power relations will be impeded or even hindered. If the starting distribution of political and economic power is unjust, the existing unjust relationships will be consolidated.

5.2 Evaluation of the GM-2: Public Interest Authority

Table 3 provides an outline of morally significant advantages and drawbacks from an access to a TWON restricted to organizations pursuing societal interests. The main difference between GM-1 and GM-2 lies in the number of societal stakeholders with access to the TWON. Whereas in GM-1 nobody is excluded, the idea of the GM-2 is to restrict access to the TWON to societal agents who pursue public interests. Thus, commercial enterprises will have no direct access to the TWON in GM-2.

Pro-Reasons GM-2	Counter-Reasons GM-2
	(R-0) GM-2 could turn out as unstable or infeasible
(B-1-1) Increase in social welfare (in an amount of ΔSW(GM-2)<ΔDSW(GM-1))	
(B-1-2) Improvement of individual quality of life (in an amount of ΔQoL(GM-2)<ΔQoL(GM-1))	(R-1-2) Risk to individual quality of life from misuse of personal data (risk lower than in GM-1)
	(R-2-1) Increased violations of the right to informational privacy
(B-2-2) Ability to prevent undermining of individual autonomy on OSNs	(R-2-2) Intensified undermining of individual autonomy
(B-2-3) Improvement of the deliberative quality of public discourse, but at least the ability to prevent the erosion of epistemic, moral, and juridical norms necessary for democratic governance	(R-2-3) Risk of collapse of institutions necessary for democratic governance
(B-3) Additional knowledge for mitigation of existing injustices (e.g., marginalization of vulnerable groups)	(R-3-1) Harms to future generations from increased resources use and/or environmental pollution (risk lower than in GM-1)
	(R-3-2) Reinforcement of existing inequalities in financial and political power relations at the national and global level

Table 3: Prima facie reasons regarding a public-interest-access to the TWON

5.2.1 Pro-Reasons (Benefits) GM-2

Welfarist Values Organizations pursuing public interests will use the TWON for improving public communication in the domains of their interest. This will increase efficiency in communication leading to economic benefits. New knowledge about complex social systems will induce innovations improving individual quality of life. Since agents with commercial interests will not have direct access to the TWON, economic benefits and improvements in quality of life will be significantly lower in GM-2 than in GM-1.

Non-Welfarist Moral Norms Jeopardized by OSNs The same prima facie reasons as for GM-1 ((B-2-2) and (B-2-3) in Table 3): TWON enables interested stakeholders in developing means for protection of democratic values and individual autonomy. In the best case, this will lead to an overall improvement of deliberative quality of online communication.

Other Moral Values and Norms The same prima facie reason as for GM-1 ((B-3) in Table 3): additional knowledge for mitigation of existing injustices.

5.2.2 Counter-Reasons (Risks) GM-2

Currently, we cannot foresee if or not this governance mode will be practically feasible. It might turn out that it is practically impossible to exclude commercial agents from TWON's access. If the latter are interested in its use, they could financially support public interest organizations with access to the TWON and commission them to use the TWON in the interests of the funding stakeholders. Or, they could commission development of a TWON for their purposes. Or, they could influence political decision makers to change the institutional setting such that they get access to the TWONs as well. The possibility that GM-2 turns out as practically infeasible provides is an additional disadvantage not reflected by the evaluative taxonomy (p. 33). In table 3 it is depicted as (R-0).

Besides R-0, the drawbacks we have identified for this governance mode are the same items as for an unrestricted access (GM-1). Merely the risk that access to TWONs will lead to misuse of personal data (R-1-2) is in GM-2 lower than in GM-1 since GM-2 provides access to a smaller number of individuals. For the same reason the risk is lower that GM-2 will harm future generations (R-3-2).

5.3 Evaluation of the GM-3: Research Community

According to the GM-3, only researchers whose research project has been approved by the TWON governing institution will be granted access to the TWON for a specific purpose (which should comply with

Pro-Reasons GM-3	Counter-Reasons GM-3
	(R-0) GM-3 turns out as unstable or infeasible
(B-1-1) Possibly increase in social welfare (B-1-2) Possibly improvement of individual quality of life	(R-1-2) Risk to individual quality of life from misuse of personal data (risk lower than in GM-2)
(B-2-2) Ability to prevent undermining of individual autonomy on OSNs (B-2-3) Improving the deliberative quality of public discourse (to a lesser extent than in GM-1 and GM-2), at least ability to prevent the erosion of epistemic, moral, and juridical norms necessary for democratic governance	(R-2-1) Increased violations of the right to informational privacy
(B-3) Additional knowledge for mitigation of existing injustices (e.g., marginalization of vulnerable groups)	(R-3-2) Risk that existing global epistemic inequalities will be increased

Table 4: Prima facie reasons regarding a TWON restricted to researchers

the research goals set by a legislative body). Table 4 offers an outline of resulting pro- and counter prima facie reasons directed at this governance mode.

5.3.1 Pro-Reasons (Benefits) GM-3

Welfarist Values Contrary to the first two governance modes, no stakeholders with direct economic interests will have access to the TWON. Still, aggregated economic benefits and improvement in quality of life are possible if a society decides to direct research activities on TWON at creation of economic benefit and improvement of quality of life. Still, the size of economic benefits to be expected from this governance mode will be lower than economic benefits expected from GM-2 (and, even more, from GM-1).

Non-Welfarist Moral Norms Jeopardized by OSNs A society which is interested in protecting norms jeopardized by OSNs – such as certain epistemic standards needed for deliberative democracy, individual autonomy – will commission research on TWON which aims at identification of mechanisms by

which the norms of interest are jeopardized and suggestion of ways to protect them. Compared to the less restricted governance modes (GM-1 and GM-2), there will be a lower number of individuals who might address an issue of interest in this governance mode. Within GM-1 and GM-2, stakeholders outside of academia can use TWON to address societal challenges from OSNs whereas, within GM-3, only authorized researchers obtain access to TWON. Accordingly, the probability that the requested mechanisms will be found, is lower in GM-3 than in GM-2 and GM-1. Similarly, due to a lower number of individuals with an access to TWONs in GM-3 compared to GM-1 and GM-2, the extent is lower to which deliberative quality of public discourses could be improved in the best case.

Other Moral Values and Norms Similarly, if a society declares topics of existing injustices as a research topic for TWON's deployment, the topic will be addressed and new knowledge on it will be generated which then might lead to more effective approaches in mitigating these injustices. Compared to the less restrictive governance modes (GM-1 and GM-2), the chances of new discoveries are lower since a lower number of individuals will have access to TWONs.

5.3.2 Counter-Reasons (Risks) GM-3

Infeasibility Similar as for GM-2, it is possible that the institutional setting of GM-3 turns out as not practically feasible. Financially or politically powerful stakeholders with an interest in using TWONs for their purposes have an incentive to get access to this technology despite institutional regulations. They might influence researchers to conduct research on TWON for their purposes, they might commission development of their own instances of TWONs or they might influence decision-makers to change the institutional setting such that they obtain access to the technology. Instability of the governance mode results additionally from political dynamics. It might turn out that a political majority comes to power which pursues public goals which oppose the ideals of current democratic societies. This majority will have the opportunity to deploy the TWON for realization of non-democratic values.

Risks for Welfarist Values According to this governance mode, a large number of individuals – potentially all researchers of complex social systems – might obtain access to a technology which can be used for elicitation of personal data. For this reason, the probability of misuse of personal data is higher than in a world without TWON, and accordingly, the risk from the TWON. However, this risk within GM-3 should be much lower than within GM-2 and GM-1.

Non-Welfarist Moral Norms Jeopardized by OSNs As discussed in Section 5.1.2 (p. 36), development and deployment of a TWON might violate the right to informational self-determination depending

Pro-Reasons GM-4	Counter-Reasons GM-4
	(R-0) GM-4 turns out as practically infeasible
	(R-2-1) Increased violations of the right to informational privacy
(B-2-3) Ability to prevent and prosecute violations of legal norms on OSNs	

Table 5: Prima facie reasons regarding a TWON governed by a judicial authority

on where the personal data for the model come from. This violation constitutes a prima facie objection to using the TWON according to this governance mode, as well.

Due to a regulated access of the TWON including specification of the purpose for which TWON is used, direct use of this technology for developing instruments for manipulation of opinions is excluded. Still, if the interest of societal stakeholders for instruments for directed manipulation of public opinion is strong enough, there will be attempts to circumvent the regulating standards, for instance by pretending to conduct research in compliance with the rules. Thus, a risk of misuse of the governing mechanism remains. How grave this risk is, depends on how the TWON governing institution will be endowed with personal and financial resources to prevent misuse. We assume a scenario in which the governing institution is sufficiently endowed to minimize the risk of misuse.

Other Moral Values and Norms Since access to the TWON is strongly restricted according to this governance mode, the possibility that the TWON will be used to preserve existing power relations remains abstract: we cannot describe a reasonable scenario how that would happen. Still, if the TWON will be developed and used in a single part of the world, global inequalities in access to knowledge and, as a result, economic inequalities might rise, if no measures mitigating a globally unequal access to the technology will be met.

5.4 Evaluation of the GM-4: Judicial Authority

The last governance mode we discuss is the most strictly regulated. Accordingly, it leads to the lowest societal benefits from TWONs but involves the lowest risks, as well. Table 5 summarizes the identified pro- and counter-reasons.

5.4.1 Pro-Reasons (Benefits) GM-4

Within this governance mode, no direct economic benefits or improvement in quality of life should be expected. It serves the only goal to enforce laws on OSN. Thereby, a society obtains an ability to prevent and prosecute violations of laws on an OSN.

5.4.2 Counter-Reasons (Risks) GM-4

Infeasibility It is possible that GM-4 turns out as infeasible or unstable as well. If it will be possible for financially and/or politically powerful stakeholders to commission development of own instances of TWONs, restriction of TWON's use to a certain societal group becomes practically infeasible. Additionally, changes in the institutional settings are possible once the governance mode has been established.

Non-Welfarist Moral Norms Jeopardized by GM-4 This governance mode does not avoid violation of the right to informational self-determination if personal data for TWON's development and use have been collected inappropriately. As for all governance modes, the possibility of misuse of the institutional setting remains for this governance mode as well. For instance, the judicial authority might mistakenly provide a too wide access to the TWON with negative consequences for the legal system. The severity of this risk depends on how the governance mode is protected against misuse and misbehaviour. Here, we assume that it is sufficiently protected to reduce these risks to a tolerable level.

6 Step 4: Exemplification of the Methodology for Identification of Ethical Commitments

Having identified pro- and counter-reasons for each of the governing modes, the question arises as to how the pro- and counter-reasons determine an overall evaluation of a governance mode. To find that out, we reconstruct justifications for and objections to the four governance modes considered in this report as deductively valid arguments (c.f. Section 2). The reconstructed arguments make explicit all the premises from which justifications and objections follow deductively. This, in turn, enables us to evaluate the identified premises in terms of their epistemic strength (plausibility, credibility).

Here, we illustrate this method by reconstructing the formal structure of the general pro- and counter-arguments directed at the claim that TWONs should (or should not) be regulated according to a governance mode i. The detailed reconstructions of the arguments for all four governance modes are in the Appendix (Section 8).

6.1 Structure of Arguments

The prima facie reasons for and against the different governance modes of the TWON reveal that the justification of a particular governance mode involves a balancing among the prima facie reasons. There are two general strategies by which a practical conclusion about a governance mode can be justified in the light of a plurality of prima facie pro- and con-reasons:

- **Debunking Strategy**: A proponent of a GM-i argues that some or all prima-facie objections to GM-i do not hold or are insignificant. An opponent to a GM-i argues that all the prima-facie reasons supporting the GM-i do not hold or are insignificant.
- **Balancing Strategy**: A proponent of a GM-i argues that the benefits from the GM-i realize such an important moral value that it is worthwhile to accept the drawbacks accompanied with the GM-i (potential benefits clearly outweigh the risks).

  For the opponents of a GM-i, we have identified two argumentative strategies. According to the first, the opponent argues that the risks of GM-i exceed certain thresholds of what is morally acceptable, presupposing a version of the precautionary principle thereby. The goal of the second counter-strategy is to demonstrate that the risks of GM-i need not to be accepted because there are alternative ways of realizing the benefits of GM-i with lower risks.

These strategies involve arguments with different premises. In the following, we reconstruct these arguments in a formal way: we use formal placeholders for empirical content which differ depending on the governance mode. The reconstruction will reveal a general structure of the arguments. Additionally, it makes visible premises whose truth-value (plausibility) will be difficult to establish. They are highlighted in violet.

Debunking Arguments

Debunking Pro-Argument

(1) [Moral Benefits from GM-i]: It is reasonable to expect that access to the TWON according to GM-i will bring about morally significant benefits B.

(2) [Moral Drawbacks from GM-i are Insignificant]: It is reasonable to expect that access to the TWON according to GM-i will lead to moral drawbacks in a morally insignificant amount.

(3) [Principle] A means M should be used if the following conditions are met:
(i) it is reasonable to expect that use of M brings about morally significant benefits B and
(ii) it is reasonable to expect that the moral drawbacks from M's use are insignificant and

(iii) for all other available means M_{alt} for which it is reasonable to expect that they realize at least B it is also reasonable to expect that they bring about moral drawbacks in a significant amount.[8]

(4) [No-Alternatives to GM-i]: According to the best available knowledge, all other means by which the benefits B could be realized involve significant moral drawbacks.

---

(5) [Conclusion]: Access to the TWON should regulated according to GM-i.

Debunking Counter-Argument

(1) [Benefits from GM-i not Morally Significant] It is reasonable to expect that access to the TWON according to GM-i will bring about benefits B for which it holds: they do not qualify as morally significant.

(2) [Moral Drawbacks from GM-i]: It is reasonable to expect that an access to the TWON according to GM-i will violate moral rights in a morally significant amount.

(3) [Principle] If it is reasonable to expect that using a means M will jeopardize moral rights in a morally significant amount and it is reasonable to expect that using M will not realize morally significant benefits, it is morally not allowed to use M.

---

(4) Conclusion: A society ought not to allow that the TWON is accessed according to GM-i.

Balancing Arguments

Balancing Pro-Argument Realization of important Benefits

(1) [Moral Benefits from GM-i] It is reasonable to expect that access to the TWON according to GM-i will bring about morally significant benefits B.

(2) [Moral Drawbacks from GM-i]: It is reasonable to expect that an access to the TWON according to GM-i will lead to moral drawbacks in a morally significant amount MD.

(3) [Benefits from GM-i outweigh the Risks from GM-i]: Benefits B have such a high moral importance that it holds:
(i) if there is a means M for which it is reasonable to expect that it realizes B and
(ii) that it involves moral drawbacks in an amount of MD and
(iii) for all other available means M_{alt} for which it is reasonable to expect that they lead at least to B it is reasonable to expect that their moral drawbacks will be higher than MD,
then M ought to be realized.

(4) [No Alternatives to GM-i]: According to the best available knowledge, all other alternatives to a TWON regulated according to GM-i by which the benefits B could be realized involve higher moral drawbacks than MD.

---

(5) [Conclusion]: Access to the TWON should regulated according to GM-i.

[8] This version of a means-end-argument leaves open how to decide if there is an alternative means M_{alt} which realizes much higher benefits B++ but involves significant moral drawbacks as well. (Thanks to Sjoerd Stolwijk for pointing this out!) If one argues that such an alternative M_{alt} should be preferred to M, then the principle in premise (3) is wrong: the specified conditions are not sufficient for the choice of the means. However, one could also argue that M_{alt} and M are both morally obliged. Then principle (3) is not wrong. However, one needs an additional explanation about how to choose between M and M_{alt}. One faces a moral dilemma with this choice: a choice which unavoidably causes moral harms. Justification of a dilemmatic choice needs an additional reasoning.
We leave this problem for the principle open. For the picture of justifications of TWON's governing modes we are depicting in this report it is not relevant.

Balancing Counter-Argument-1 Violation of Moral Thresholds

(1 [Threshold Precautionary Principle]: If it is seriously possible that a use of a means M leads to violation of moral rights beyond a threshold of morally acceptable R_{min} then a society ought not to allow that M is realized independently of the expected benefits from M.

(2) [Moral Drawbacks from GM-i]: It is seriously possible that an access to the TWON according to GM-i will violate moral rights in a morally significant amount $R(GM_i)$.

(3) Violation of moral rights in an amount $R(GM_i)$ transgresses the threshold of morally acceptable R_{min}.

---

(4) [Conclusion]: A society ought not to allow that the TWON is accessed according to GM-i.

Balancing Counter-Argument-2 Availability of Alternatives

(1) There is a prima facie moral requirement that B.

(2) From all available means $M_1, ..., M_n$ for which it is reasonable to expect that their realization will lead to B, only the means M_i ought to be chosen for which it is reasonable to expect that its realization has the lowest moral drawbacks.

(3) [Benefits and Drawbacks from GM-i]: It is reasonable to expect that an access to the TWON according to GM-i leads to B and that it will violate moral rights in a morally significant amount R_i.

(4) [Alternatives to GM-i with lower moral costs available]: It is reasonable to expect that M_a realizes B and that it will violate moral rights in an amount $R_a < R_i$.

---

(5) [Preliminary Conclusion from (3)&(4)]: It is not the case that it is reasonable to expect that an access to the TWON according to GM-i has the lowest moral drawbacks for attaining B.

---

(6) [Conclusion from 1&2&5]: A society ought not to allow that the TWON is accessed according to GM-i.

Clarifications of the Formal Reconstruction The arguments are reconstructed in an **epistemic form** ("it is reasonable to expect that p") instead of an assertoric form (p). This makes the premises with empirical content (premises (1), (2), and (4) in the Pro-Argument and the premises (1) and (2) in the Counter-Argument) justifiable. However, the formal principles (Premises (3) in both arguments) are wrong in the epistemic form. The conditionals in the premises (3) are not true: since we as humans are fallible in our abilities to forecast, our expectations are sometimes wrong. In those cases it might be reasonable to expect a benefit from a certain means and to believe that the means should be undertaken, but, actually, the means does not realize the reasonably expected outcome. In that case it does not hold that the means should be undertaken. With the epistemic reconstruction of the argument, we accept this drawback of the argument, for it better represents the way how we reason (we do rely on formal principles which are wrong in a strict sense). Reconstructing the arguments in an assertoric form would bring us similar difficulties with premises with empirical content (c.f. Section 2 for additional explanations).

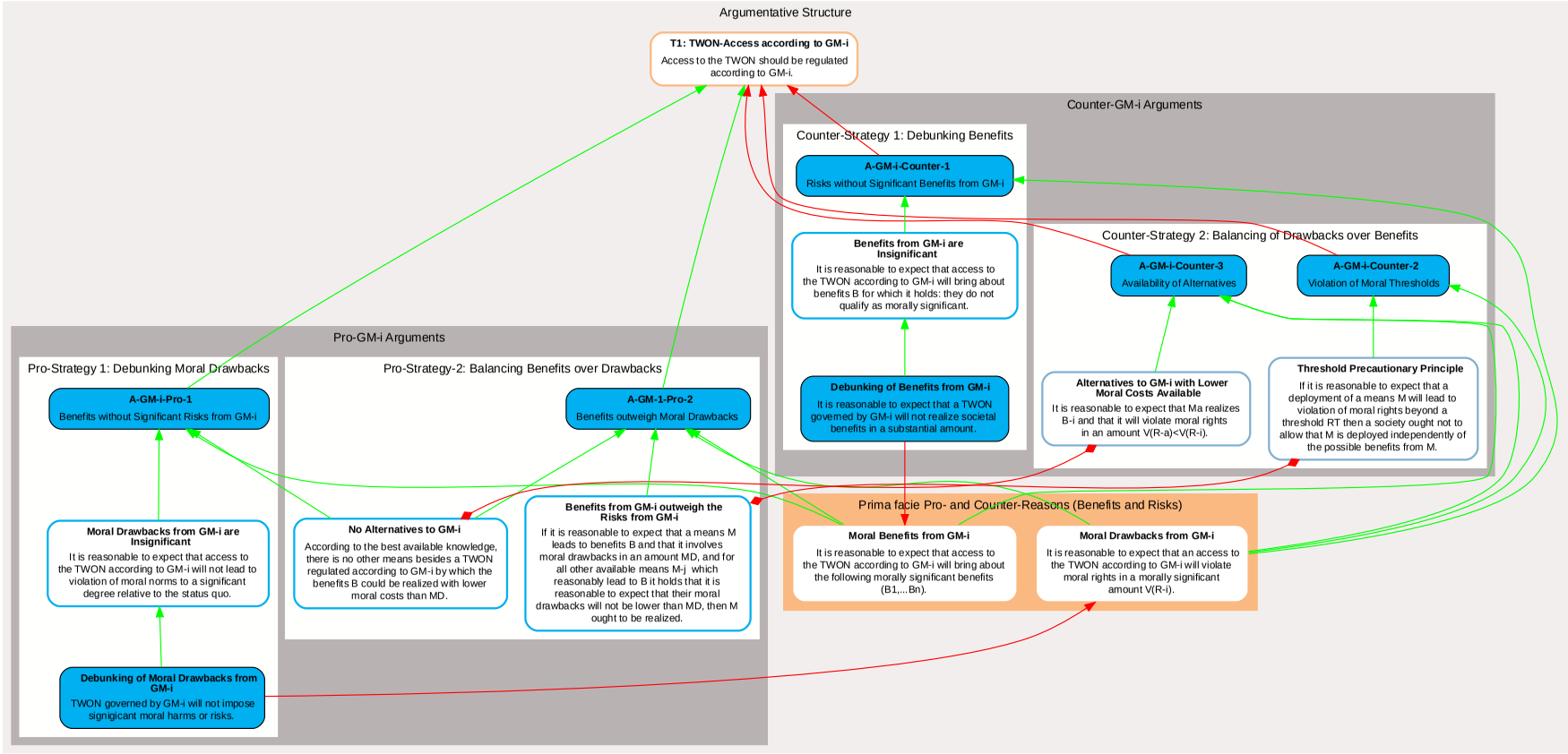

Figure 5: Formal Structure of the Dialectics on TWON-deployment

The reconstruction contains several **unspecified concepts**, e.g., ”morally significant”, ”moral norms”, ”significant degree of violation of a moral norm”, ”reasonable to expect”, ”seriously possible”. These first three terms articulate a moral evaluation, the last two an epistemic evaluation. The content of these terms depends on the underlying moral and epistemic theory. Since we do not presuppose a specific moral or epistemic theory, we use these terms abstractly:

- an outcome o is **morally significant** if and only if o has an impact on what the background moral theory considers as morally considerable (for instance, according to utilitarian theories, all actions which change the amount or the distribution of utility in the world, are morally significant).
- the set of moral norms is, similarly, determined by the background moral theory. In Section 5 on page 33 we have outlined a general taxonomy of moral norms which we have presupposed for this report.
- the notions ”**reasonable to expect** that p” and ”**seriously possible** that p” articulate a qualitative evaluation of credibility that a possible state of the world p will be realized. In this report, both terms express the following idea: it is consistent with the available background knowledge that the possible state of the world p will be the case and there are further, plausible, explanations how p would turn out to be the case (these explanations provide reasons for the expectation that p or qualify the possibility that p as a serious one).

For the sake of demonstrating the argumentative structure, an unspecified use of the notions should suffice. When it will come to deriving practical conclusions, these concepts must be specified.

6.2 Discussion of the Formal Reconstruction

Figure 5 represents the dialectical structure of the arguments. The orange box represents the starting statements for the argument analysis: it contains the two statements justified in the previous section that a TWON will bring about certain benefits and risks. At the top of the figure, the normative statement of interest is depicted by a white box with an orange line. It is supported by two arguments illustrated in the left gray box ”Pro-GM-i-Arguments” and attacked by three arguments depicted in the right gray box ”Counter-GM-i-Arguments”. The white boxes with the blue lines represent the controversial premises of these arguments.

The map illustrates that the debunking-arguments attack the assumptions of the analysis, i.e. the statements that TWONs realize benefits and that they involve risks (red arrows from <Debunking of Moral Drawbacks from GM-i> to [Moral Drawbacks from GM-i] and from <Debunking of Benefits from

GM-i> to [Moral Benefits from GM-i]). Accordingly, the plausibility of the debunking-arguments depends on whether these attacks can be reasonably justified. The deductively valid reconstruction of the arguments reveals which premises are crucial here: premise (2) of the Debunking-Pro-Argument and premise (1) of the Debunking-Counter-Argument. These premises are highlighted violet in the reconstructions above.

Contrary to the debunking arguments, the balancing arguments accept the prima facie pro- and counter-reasons identified in the previous sections: premises (1) and (2) in the Pro-Balancing-Argument, premise (2) in the threshold-counter-argument; and premises (1) and (3) in the no-alternatives-counter-argument. These empirically informed premises are uncontroversial among the proponents of the opposing positions. However, as the map in Figure 5 depicts, balancing-arguments attack each other.

Firstly, there is a contradiction between the statement [Benefits of GM-i outweigh the Risks from GM-i] and the statement [Threshold Precautionary Principle]. These statements are general principles of the Balancing Pro-Argument (premise (3) of <A-GM-i-Pro-2>) and the Balancing Counter-Argument-1 (premise (1) of <A-GM-i-Counter-2>). Premise (3) in the Pro-Balancing argument is a version of a risk-acceptance principle for it claims that an action which leads to certain important benefits should be taken despite it risks. Premise (1) of the Balancing Counter-Argument-1 is a version of a risk-avoiding principle, the precautionary principle. The general principles of the Balancing-Arguments are therefore controversial, we have highlighted them as violet in the reconstruction.

Secondly, a contradiction is depicted in the map (Figure 5 between the statement [No Alternatives to GM-i] (which is a premise of both pro-arguments, premise (4)) and the statement [Alternatives to GM-i with lower moral costs available] which is a premise of the Balancing-Counter-Argument-2 (premise (4) of <A-GM-i-Counter-3>). That the two pro-arguments and the third counter-argument presuppose an premise about the availability of alternatives to GM-i is a result of the deductively valid reconstruction of the arguments. The dialectical structure reveals that these premises are controversial.

Summary 3

We can learn the following points about the burdens of justification of particular governance modes of TWONs already from the formal structure of pro- and counter-arguments:

- **Debunking** pro- and counter-arguments face the challenge of justification of their claim that the expected benefits and risks from TWONs identified so far will not be realized.
- **Balancing** pro- and counter-arguments rest on general principles which are, strictly speaking, false. Thus, the burden of justification lies on making these principles as plausible as possible.
- An additional controversy between the supporters of the pro- and counter-arguments is directed at the **availability of reasonable alternatives** to the governance mode in question.

In the following section, we jump to our interpretation of outcomes of the detailed reconstruction of arguments. The reconstructed arguments and the resulting argument maps are documented in the Appendix (p. 60).

7 Discussion of the Results from the Argument Analysis

7.1 Lessons from the Argument Reconstruction

This report is motivated by the fact that TWON is a technology which – should its development succeed – offers tremendous societal benefits on the one hand but involves grave societal risks on the other. TWONs aim to provide orientation on how to govern large online social networks which jeopardize democratic institutions. However, their deployment can exacerbate societal challenges it aims to mitigate. Among the various ethical questions raised by such a technology, this report has analysed the questions of how research and deployment of TWONs can be reasonably justified and what the implications are for the research process in light of its potential benefits and risks.

In this report, we have firstly presented a methodology by which these questions can be addressed (Section 2). According to it, justifications of the ethical claims of interest – in our case a recommendation to use the technology and its negation – are reconstructed as deductively valid arguments. The reconstruction allows an epistemic evaluation of the revealed premises (in terms of an epistemic value such as plausibility/ credibility/ reasonableness/ truth-value).

Secondly, we have conducted an argumentation analysis based on the currently available foreknowledge about the expected benefits and risks from TWONs (Sections 3 to 6). For that, we have constructed four scenarios about how a society regulates access to TWONs. Subsequently, we have reconstructed arguments supporting and rejecting each of the four governance modes. The detailed arguments are documented in the Appendix (Section 8 on p. 60). In the following, we summarize the main results of the argument analysis.

Lesson 1: The extent of the benefits and risks of TWONs depends on societal choices Reasons for and against the use of TWONs crucially depend on the governance mode of this technology (Section 5). Whereas an unrestricted access to TWONs promises important societal benefits but involves grave risks, a strictly regulated access diminishes societal benefits from the technology but reduces its risks as well. This implies that the motivating dilemma – how to deal with a technology which promises societal benefits and involves grave risks – boils down to the question of how a society ought to manage its benefits and risks. There is a plenitude of choices which raise their own ethical trade-offs (regarding a distribution of benefits and risks of TWONs). To avoid that a governance mode of a TWON just emerges by the fact of its technological availability, an (academic) analysis of the available governance modes and public deliberations about their ethical trade-offs are necessary.

Lesson 2: At the current state of foreknowledge, no governance mode of a TWON can be reasonably rejected Reconstruction and evaluation of arguments for and against the four governance modes (Section 3) on the basis of the currently available foreknowledge about possible outcomes (Section 4) has demonstrated that justifications of all discussed governance modes rest on arguments with uncertain premises. For all governance modes, we have constructed supporting arguments which are at least controversial and we could not construct uncontroversial and reasonable counter-arguments.

This result is not trivial: It has not turned out – contrary to our initial expectation – that some extreme governance modes can be justified only by arguments which presuppose obviously implausible premises. In other words: we could not confirm the hypothesis that the extreme governance modes are obviously unacceptable.

This result does not imply that all governance modes are morally acceptable in light of the currently available foreknowledge. A positive claim about the moral status of TWON's governance modes remains uncertain. Since there is no need to make decisions about TWON's governance immediately, resources should be spent for analysis of the identified uncertainties and public deliberation of revealed ethical trade-offs.

Lesson 3: TWON might turn out as a technology whose use should be thoroughly regulated At the current state of knowledge, no governance mode is morally unacceptable and it remains unclear which of them are within the realm of morally acceptable. The dialectical structure of the detailed reconstructions has revealed that justification of the unrestricted mode of governance GM-1 crucially depends on the claim that the stricter regulated modes of governance (GM-2 to GM-4) are practically infeasible (c.f. Figure 8 and discussion on pages 62 ff.). If it turns out that TWON's regulation is possible, justification of GM-1 is committed to the claim that the risks of TWONs will not be materialized to a substantial degree. This claim faces a demanding burden of justification (though, we do not consider this claim to be obviously implausible, see our discussion in Section 8.1.3).

Thus, it might turn out after a more detailed inquiry into risks of TWONs that an unrestricted access to TWONs is morally unacceptable. This, in turn, would have substantial implications for the research process (we address it in the next subsection).

Lesson 4: Recommendations for Additional Inquiries The outcomes of the argument reconstruction depend crucially on the quality of the available foreknowledge. Its quality changes with development of the technology. Accordingly, the empirical content of the premises of the reconstructed arguments will change. That means that argument reconstruction justifies recommendations relative to a certain body of foreknowledge which changes in time. Since there is no need to make decisions about TWON's use at this point in time, the argument reconstruction can be used for insights on which kind of foreknowledge is most relevant for practical reasoning.

Inquiry 1: Feasibility of regulation of access to TWONs: We have discussed three scenarios on how access to TWON could be regulated. They all assumed that it is practically possible that a governing institution excludes certain groups of societal stakeholders from access to TWONs. We pointed out the governance modes need institutional sophistication and sufficient resources to enable their functioning (c.f. sections where we described risks for misuse of the governance modes in Section 5). And we have pointed out the possibility thatthe governance modes which restrict access to TWON may turn out as practically infeasible. Since the practical infeasibility of these governance modes plays a crucial argumentative role, their more precise empirical assessment would help to take an informed position in the debate.

Inquiry 2: Alternatives for Protection of Democratic Values and Law Enforcement on OSNs: The arguments which justify the two restricted governance modes of a TWON (GM-3 and GM-4) contain premises whose truth-value we currently cannot decide. A crucial premise in all these pro-arguments is the claim that no alternative ways are available by which the intended moral goals – protection of

democratic values and laws enforcement on OSNs – can be attained with lower moral drawbacks.

This claim remains undecided because, at the current stage of research, it is unclear whether such an ambitious tool as a TWON – a digital twin of an online social network – is necessary to achieve the desired goals. The following potential objections arise:

- The Digital Service Act (DSA) provides authorised researchers (according to Article 40(8) of DSA) the right to obtain data from what the European Commission defines as ”very large online platforms” for conducting research on systemic risks (specified in Article 34(1) of DSA) that includes protection of norms necessary for democratic governance (”Negative effects on civic discourse and electoral processes”) and law enforcement (”dissemination of illegal content”). In light of this regulation, why would statistical analysis of the relevant data from OSN not suffice to identify illegal behaviour and behaviour which threatens democratic values?

- to understand emergent outcomes on existing large social networks, why are toy-models of complex social networks not sufficient? Toy-models of an OSN represent certain structural properties of an OSN (e.g., the small-world property, fat-tail degree distribution, degrees of homophily varying with different attributes) without including personalized data and without allowing reliable forecasts of existing social networks.

These topics need to be addressed to provide more informed justifications of the governance modes 3 and 4.

Inquiry 3: Description of the current situation with OSNs and TWONs: Plausibility of the arguments supporting particular modes of governance differ. The more promising argument supporting unrestricted access to TWONs is the debunking argument (p. 60). Its plausibility depends on a more detailed knowledge of the current situation with OSNs and TWONs:

- to which degree have the existing OSNs already damaged democratic institutions, individual autonomy, and the right to informational self-determination?

- are their stakeholders who have undisclosed already developed or are developing similar technologies to TWONs with the aim to influence public opinion?

These topics need to be addressed to provide more informed justifications of the governance modes 1 and 2.

Summary 4

TWON is a technology which enables societies to protect democratic governance in light of large online social networks but which can also exacerbate the challenges it aims to mitigate. What do we learn from the argument analysis about how to deal with such a technology and its research process?

The lessons from our analysis are:

- The risks and benefits of the TWON hinge upon the manner and extent to which access to this technology is regulated. There is a plenitude of options, ranging from unlimited access to a very highly controlled usage. Each mode of governance entails distinct societal societal benefits and risks.
- Given the now available foreknowledge about possible consequences of TWON's deployment, none of the governance modes discussed in this report – from an unrestricted access to an access by permission of a judicial authority only – could be excluded as obviously unacceptable.
- Should new information emerge, regulating the usage of TWONs could prove to be necessary.
- Argument analysis has revealed premises in the justifications of different governance modes that we deem uncertain. These premises need to be addressed by future research or thorough deliberation before an informed decision about the deployment of TWONs is possible.

7.2 Implications for the Research Process

The argument reconstruction has revealed ethical controversies that will emerge when it will come to deciding how to use TWONs. In this section, we turn to the question of what these results imply for the process of developing of the technology.

Implications from Lesson-3: TWON might turn out as a technology which should be thoroughly regulated If an in-depth analysis of various governance modes of TWONs comes to the conclusion that TWONs access ought to be strictly regulated since there are good reasons not to allow that it is publicly accessible, this outcome will have implications for the research process: research outcome

which would enable third parties to build a TWON should not be publicly accessible as well. This, in turn, would require the research process to be organized appropriately.

These implications affect both, the immediate research process and research politics.

- Based on these implications, we recommend for the research process:
  - Ex-post evaluation: Installation of a regular assessment of the project's advancement which specifies if the research process has reached a stage in which its outcomes should not be publicly accessible. This could be integrated into project meetings. Discussion of the research outcomes – publications and the platform model – at project meetings regarding the question if the recent insights (e.g., insights on OSN's algorithms or on effects from platform design choices on opinion formation) and the current stage of the platform model can be used for purposes which are objectionable on moral reasons.
  - Ex-ante evaluation: The project meetings should also be used to assess the morally relevant risks from the intended outcomes of the next step of the research process.
  - Discussion of criteria for the evaluation of the practical relevance of the research process.
- For the research politics, we see the following task:
  - A way to organize and to fund public research on TWONs needs to be specified whose outcome should not be freely accessible should it turn out that access to TWONs ought to be restricted.

Implications from Lesson-4: Controversy on Alternatives A crucial goal of TWON's development is creation of an instrument for analysis of causal effects in a large online social network (c.f. Section 1). For instance, the following question could be addressed on a TWON: Do the designed properties (algorithms for the distribution of content) of an online social network cause problematic communicative outcomes (polarization, spread of mis- and disinformation) to a degree which, in certain circumstances, inflame political riots?

The argument which supports TWON-development contains the claim that causally contributing factors of outcomes in such a complex system can be identified only by availability of a perfect digital copy of that system such that single properties of the algorithms can be systematically varied. The argument which attacks TWON-development contains the contrary claim: there are other means by which properties of the algorithms can be identified which causally influence (inflate, reinforce etc.) problematic outcomes.

Both claims are unjustified: To justify the supporting claim, one needs to justify a negative claim that there are no alternative means for the causal analysis. It is not clear how such a negative claim can be justified. The contrary claim can be justified by an example of an alternative means for the targeted causal analysis. Such an alternative is currently not known. But it might be the case that there are some (they have not been discovered yet). For the following options cannot be excluded either: With sufficient data from an OSN it might be possible by statistical analysis to identify causally influential properties of the distribution algorithm. Or, more simplified models of an OSN than a TWON might be sufficient for causal analysis.

Thus, the pro as well as the counter-argument rest on uncertain statements. What follows from that for the research process?

The uncertainty arises because it is unknown how precise a model of an OSN needs to be to provide means for analysis of causal effects of OSN's distribution algorithms. To put it differently: what is the most distorted model of an OSN which allows analysis of causal effects of OSN's algorithms? This question can be approached by the means of modelling.

Testing by the means of Fictional Objectivity The main idea is the following: we create a set of fictional networks which replicate the structural properties of the target OSN – set of agents with a set of characteristics which are connected within the network according to the network structure of the target OSN and which interact with each other (produce content or react to it) according to the interaction structure of the target OSN. The created fictional networks might differ in the distribution of agent's characteristics or in behaviour rules or in the algorithms which distribute content to the agents. Thereby we create a fictional plurality of possible online social networks which with the same structure as the target OSN. The created fictional OSNs (F-OSN) serve as target systems for systematic testing – therefore they provide the fictional objectivity. We create distorted models of the F-OSNs (Model-F-OSN) which deviate from the F-OSN in certain properties of interest. We test in how far the outcomes of distorted models of F-OSN differ from the fictional reality realizing F-OSNs.

If the differences can be explained by standard statistic models, there is no evidence that the properties of interest cause emergent effects. Then, they do not need to be precisely represented in the TWON of the target OSN. Their value can be approximated or distorted such as to anonymize it. If distortions in Model-F-OSNs cause emergent effects relative to the F-OSN, the respective property should be represented realistically in the TWON. Figure 6 depicts the idea of how fictional modeling can help build a minimal TWON for causal analysis.

Creation of fictional social networks on which particular properties are tested for their dynamic ef-

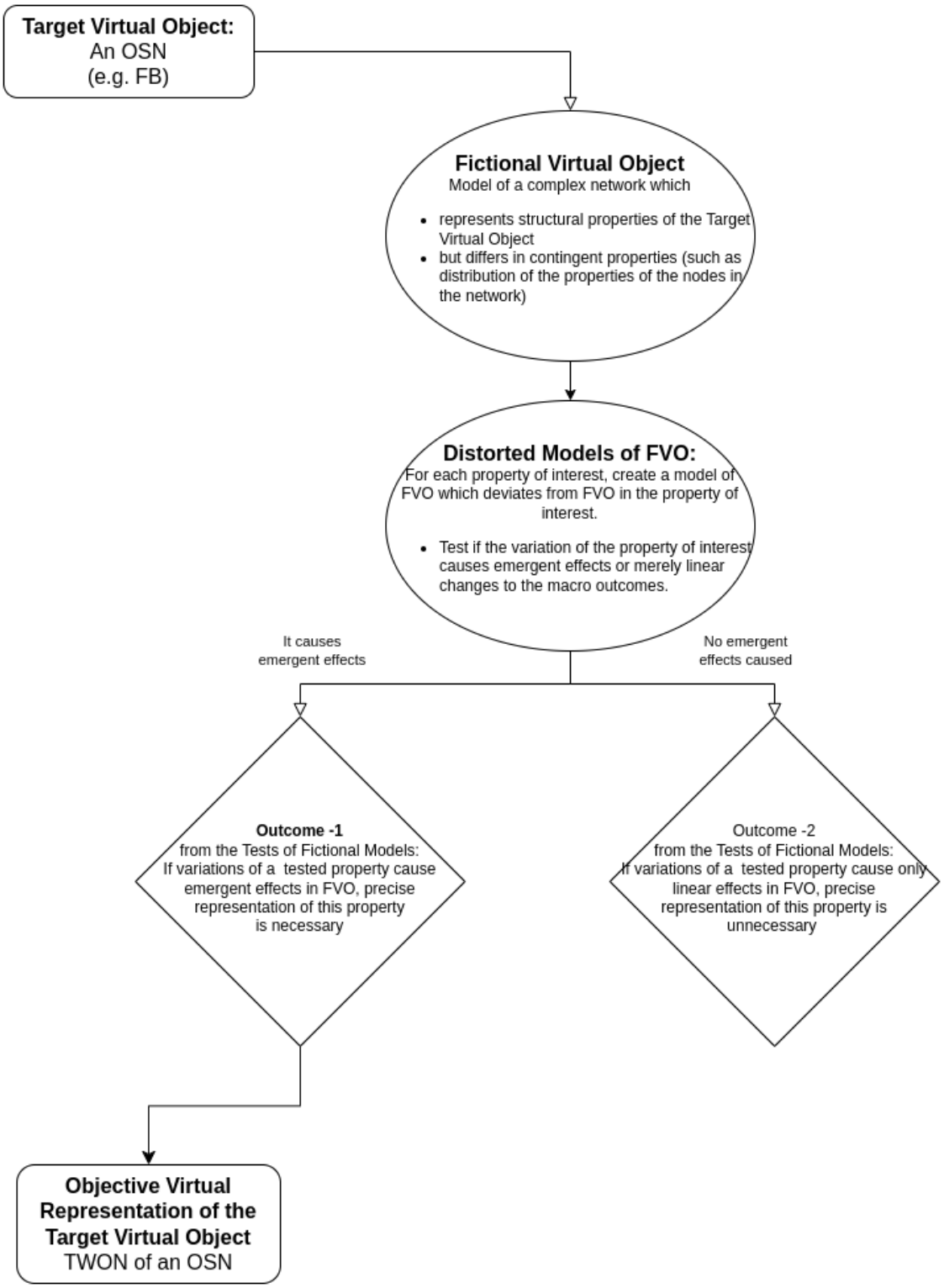

Figure 6: Modelling of Fictional Objectivity

fects is the way by which a justification for the necessary degree of similarity between the TWON and the target OSN can be provided. Implementation of this testing by the means of fictional objectivity is therefore a way how to proceed within the research process of a TWON in light of uncertainty about alternatives for causal analysis of OSN.

Personal Data Usage Another motivating point for the ethical analysis is the insight that TWON's functioning requires detailed users' data from the target OSN (c.f. Section 1.3). This raises the question of how to deal with the trade-off between TWON's functionality and the right to informational privacy.

We have described three scenarios for origins of personal data used for TWON's development and operation (c.f. Section 4.1). In our evaluation of their ethical significance (Section 5.1.2 on 36) we concluded that the scenario in which personal data have been acquired in compliance with the right to informational self-determination is practically hardly realizable. It is rather reasonable to expect that personal data for TWON will be acquired in a way which either violates the right to personal self-determination or for which it is contested whether it contradicts this right.

We see two options for dealing with this ethical conflict. The first option is directed at researchers and TWON hosting institutions. The second option is directed at public decision-making processes.

Option 1: Identification of the necessary data Reconstruction of arguments has demonstrated that arguments justifying a certain governance mode of a TWON need to prove that TWON is not only a sufficient means for achievement of certain desired goals but that it is also necessary for that goal (this claim appeared in the no-alternative premises in the pro-arguments). In the previous section, we have suggested a way how this necessity-requirement could be satisfied in the process of development of the TWON: by testing for emergent effects on fictional-reality models (c.f. p. 56).

This reasoning applies to data usage as well: precise personal data from an OSN is sufficient for TWON's functionality. However, it is unclear which amount of personal data is necessary for a desired degree of TWON's functionality. This, in turn, could be tested by the means of fictional-reality models. If it turns out that certain personal data are not necessary for the target functionality of a TWON, the ethical conflict does not arise.

Option 2: Deliberation about the trade off between violation of privacy rights and societal benefits from TWON Still, it can turn out that personal data from an OSN which have been acquired without informed consent of the users is necessary for TWON's functionality at a targeted level. That would imply that an ethical trade-off is unavoidable. This needs to be addressed in deliberative processes involving the relevant public outside of the research community.

7.3 Monitoring Strategy

To monitor our recommendations for the research process, we suggest:

- For internal reviews of project's deliverables: include the following ex-post evaluation points for the reviewer:
  - Can the research outcomes presented in the deliverable be used for purposes which are morally objectionable?
  - If the research outcome has been based on sensible personal data, is there evidence that the use of this kind of personal data is necessary for the intended research outcomes?
- For project meetings: include an ex-ante monitoring of potential ethical risks for the then following research steps. For that, similar questions can be addressed when planning the subsequent research:
  - Can the expected outcomes from the next research step turn out to be usable for purposes which are morally objectionable?
  - If sensible personal data is intended to be used in the following research step, has an analysis of the question if this personal data is necessary for the intended outcome been conducted or been included into the research planning?

Summary 5

Regarding the research process, our analysis suggests:

- In case TWONs turn out to be a technology requiring strict regulation, then the research and development process must also be subject to regulatory oversight.
- 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 conclude that such analyses be conducted.

8 Appendix: Detailed Reconstruction of Arguments

8.1 GM-1: Unrestricted Access

In this section, we present reconstructed arguments directed at the claim [T1-GM-1: Access to the TWON ought to be unrestricted.] and its negation by specifying the premises from formally reconstructed arguments in Section 6.1. Premises which we consider to be "implausible" are highlighted in red. Premises which we consider to be "uncertain" are highlighted in violet. In Section 8.1.3, we justify our evaluations.

8.1.1 Arguments Supporting GM-1

Argument-GM-1-Pro-1: Debunking of Counter-Reasons

(1) [Moral Benefits from GM-1]: It is reasonable to expect that an unrestricted access to the TWON will bring about the following morally significant benefits:

(B-1-1) it enhances social welfare (in an amount of Δ SW(GM-1));

(B-1-2) it improves individual quality of life (in an amount of Δ QoL(GM-1));

(B-2-1) it enables that undermining of societal norms necessary for democratic governance is prevented;

(B-2-2) it enables that undermining of individual autonomy on OSN is prevented;

(B-3) it will contribute to a mitigation of existing injustices.

(2) [Moral Drawbacks from GM-1 are Insignificant]: It is reasonable to expect that unrestricted access to the TWON will lead to violations of moral rights in a morally insignificant amount relative to the status quo.

(3) [Principle] A means M should be used if the following conditions are met:
(i) it is reasonable to expect that use of M brings about morally significant benefits B and
(ii) it is reasonable to expect that the moral drawbacks from M's use are insignificant and
(iii) for all other available means M_{alt} for which it is reasonable to expect that they realize at least B it is also reasonable to expect that they bring about moral drawbacks in a significant amount.

(4) [No-Alternatives to GM-1 without Drawbacks]: According to the best available knowledge, there is no other means by which the benefits (B-1-1,...,B-3) could be realized without significant moral drawbacks.

---

(5) [Conclusion]: Access to the TWON ought to be unrestricted.

Argument-GM-1-Pro-2: Balancing Benefits over Risks

(1) [Moral Benefits from GM-1] It is reasonable to expect that an unrestricted access to the TWON will bring about the following morally significant benefits:

(B-1-1) it increases social welfare (in an amount of ΔSW(GM-1));

(B-1-2) it improves individual quality of life (in an amount of ΔQoL(GM-1));

(B-2-1) it enables that undermining of societal norms necessary for democratic governance is prevented;

(B-2-2) it enables that undermining of individual autonomy on OSN is prevented;

(B-3) it will contribute to a mitigation of existing injustices.

(2) [Moral Drawbacks from GM-1]: It is reasonable to expect that an unrestricted access to the TWON will involve moral drawbacks in a morally significant amount:

(R-1-1) Social welfare reduction from additional resources consumption;

(R-1-2) Risks to individual quality of life from misuse of personal data;

(R-2-1) Reinforcement of violations of the right to informational privacy;

(R-2-2) Intensification of undermining of individual autonomy;

(R-2-3) Collapse of institutions necessary for democratic governance;

(R-3) Reinforcement of existing injustices.

(3) [Protection of Democracy Trumps Moral Drawbacks]: Protection of democracy has such a high moral importance that it holds:
If there is a means M for which it is reasonable to expect that the society obtains the ability to prevent collapsing of democratic norms and whose deployment involves the risk that it leads to a collapse of democratic institutions (R-2-3) and involves further moral drawbacks R-1-1, R-1-2, R-2-1. R-2-2, R-3
and for all other available means M_{alt} for which it is reasonable to expect that they protect democratic institutions it is reasonable to expect that they will involve moral drawbacks higher than (R-1-1, R-1-2, R-2-1. R-2-2, R-3)
then M ought to be realized.

(4) [No Alternatives to GM-1 with less Drawbacks]: According to the best available knowledge, all alternative ways to protect democratic institutions involve higher moral drawbacks than (R-1-1, R-1-2, R-2-1. R-2-2, R-3).

---

(5) [Conclusion]: Access to the TWON ought to be unrestricted.

8.1.2 Arguments Against GM-1

Argument-GM-1-Counter-1: Debunking-Benefits

(1) [Benefits from GM-1 Morally Insignificant]: It is reasonable to expect that an unrestricted access to the TWON will not bring about any societal benefits in a significant amount.

(2) [Moral Drawbacks from GM-1]: It is reasonable to expect that an access to the TWON according to GM-1 will involve moral drawbacks in a morally significant amount:

(R-1-1) Social welfare reduction from additional resources consumption;

(R-1-2) Risks to individual quality of life from misuse of personal data;

(R-2-1) Reinforcement of violations of the right to informational privacy;

(R-2-2) Intensification of undermining of individual autonomy;

(R-2-3) Collapse of institutions necessary for democratic governance;

(R-3) Reinforcement of existing injustices.

(3) [Principle]: If it is reasonable to expect that a means M jeopardizes moral rights to a morally significant degree and it is reasonable to expect that M will not realize morally significant benefits, it is morally not allowed to realize M.

---

(4) Conclusion: A society ought not to allow that the TWON is accessed freely.

Argument-GM-1-Counter-2: Risks from GM-1 Transgress Moral Thresholds

(1) [Threshold Precautionary Principle]: If it is seriously possible that a use of a means M leads to violation of moral rights beyond a threshold of what is morally acceptable R_{min} then a society ought not to allow that M is realized independently of its expected benefits.

(2) [GM-1 Can Collapse Democracy]: It is seriously possible that an unrestricted access to the TWON will lead to a collapse of institutions necessary for democratic governance in the long run (R-2-3).

(3) Collapse of democratic governance is an outcome beyond a threshold of what is morally acceptable.

---

(4) Conclusion: A society ought not to allow that the TWON is accessed freely.

Argument-GM-1-Counter-3: Alternatives to GM-1

(1) There is a prima facie moral requirement that a democratic society has the ability to protect societal norms necessary for democratic governance.

(2) From all available means $M_1, ..., M_n$ for which it is reasonable to expect that they protect democratic norms, only the means M_i ought to be chosen for which it is reasonable to expect that its realization has the lowest moral drawbacks.

(3) [Benefits and Drawbacks from GM-1]: It is reasonable to expect that an unrestricted access to TWONs (i) will protect democratic institutions (B-2-1), (ii) will realize further benefits (B-1-1, B-1-2, B-2-2, B-3) and (iii) that it involves moral drawbacks (R-1-1,...,R-3).

(4) [Alternatives for Democracy Protection with Lower Moral Drawbacks Available]: There are means for which it is reasonable to expect that they protect democratic norms and involve moral drawbacks lower than (R-1-1,...,R-3):

<+ <Alternatives to GM-1 with Lower Moral Drawbacks>: GM-3 is expected to protect democratic values and involves only the moral drawbacks (R-2-1) and (R-3).

---

(5) [Preliminary Conclusion from (3)&(4)]: It is not the case that it is reasonable to expect that an unrestricted access to the TWON protects democracy with the lowest moral drawbacks.

---

(6) [Conclusion from 1&2&5]: A society ought not to allow that the TWON is accessed freely.

Figure 7 depicts the resulting dialectical structure.

8.1.3 Discussion

GM-1 supporting arguments The Debunking-Pro-Argument (<A-GM-1-Pro-1>) contains the claim that an unrestricted access to the TWON does not involve significant moral risks (premise (2)). This claim contradicts the outcome of identification of risks from TWON and implies that all the risks identified in Section 5 will most likely not materialize. Thereby, a proponent of the Debunking-Pro-Argument is committed to the following claims:

- it is unlikely that an unrestricted access to the TWON will reduce social welfare or individual quality of life or welfare losses will be neglectable;

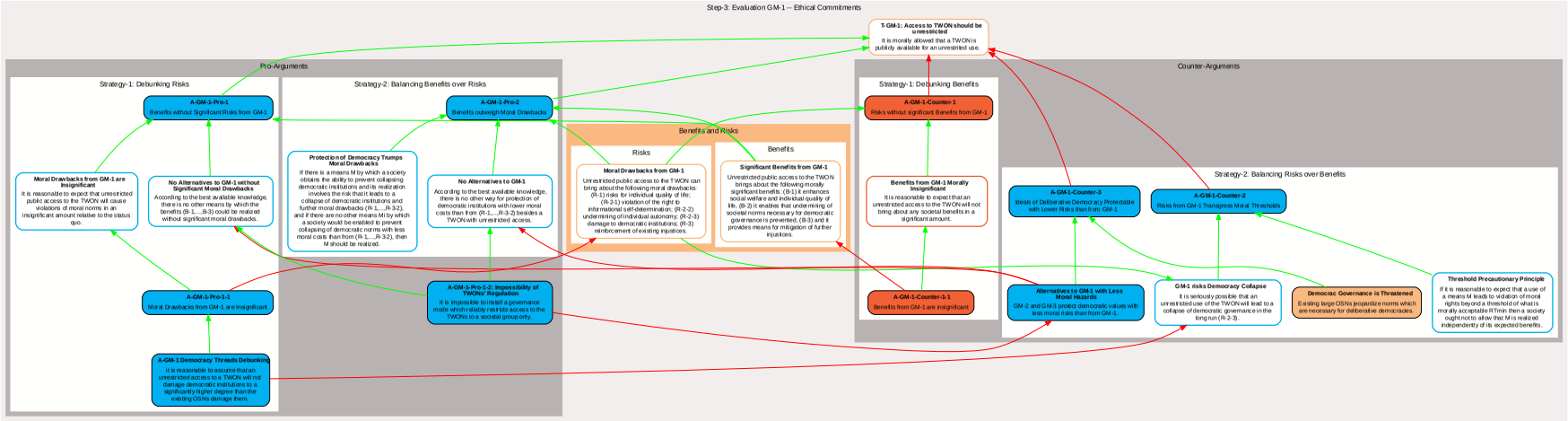

Figure 7: Dialectics about GM-1

Arguments and statements in orange represent starting assumptions of the argument analysis: the two statements (justified in Section 5.1) that a TWON according to GM-1 will bring about certain benefits and risks. At the top of the figure, the normative statement of interest [T-GM-1] is depicted by a white box with an orange line. Additionally, the argument <Democratic Governance is Threatened> which justifies a premise in the balancing-argument <A-GM-1-Counter-3> has been implicitly assumed at the outset of this study (c.f. Section 1 and references there for its elaboration).

The main statement [T-GM-1] is supported by two arguments depicted in the gray box ”Pro-Arguments” on the left and and attacked by three arguments depicted in the gray box ”Counter-Arguments” on the right.

Arguments and statements which we have evaluated as ”implausible” are coloured red; arguments and statements in blue we have evaluated as ”uncertain”.

- It is unlikely that unrestricted access to the TWON will violate the right to informational self-determination or the right to informational self-determination does not hold;
- it is unlikely that an unrestricted access to the TWON will undermine individual autonomy or an unrestricted access to the TWON will undermine individual autonomy to an insignificantly higher degree than it is undermined by existing OSNs;
- it is unlikely that an unrestricted access to the TWON will reinforce threads for democratic governance or an unrestricted access to the TWON will damage democratic institutions to an insignificantly higher degree than the existing OSNs damage them.
- it is unlikely that an unrestricted access to the TWON will reinforce existing injustices or an unrestricted access to the TWON will reinforce existing injustices to an insignificantly higher degree than the existing OSNs reinforce existing injustices.

These claims require a profound justification. Currently, we do not see how they can be reasonably justified. But this does not imply that no justification exists. For instance, it can be the case that the existing OSNs have already damaged societal norms and practices necessary for deliberative democracy, that they are already violating the right to informational privacy and individual autonomy to such degree that a free access to a TWON will not cause additional damage in a significant amount. Or, it might be the case that some (epistemically and financially endowed) societal stakeholders have already developed or started developing TWONs and that only a public availability of a TWON will enable a society to ensure that some agents do not manipulate public opinions in their interests.

Additional research is needed to evaluate the plausibility of this reasoning. In Figure 7, a justification consisting of two arguments (<A-GM-1-Pro-1-1> and <A-GM-1 Democracy Threads Debunking>) is depicted. It reflects the idea that an unrestricted access to TWONs will not add additional threads to democratic governance and other endangered moral values besides their undermining by existing OSNs. Currently, we cannot judge how plausible this reasoning is. For this reason, we evaluate the premise (2) and the arguments depicted as its support as uncertain.

Additionally, premise (4) in the Debunking Pro-Argument is uncertain. It claims that no alternative means are available by which the same benefits will be achieved without moral drawbacks in a significant amount. As we will discuss in the next sections, the most important outcomes will be achieved with the more restricted governance modes (GM-2 and GM-3) as well. Moreover, there is a further research attempt aiming at developing instruments for analysis of emergent outcomes on OSNs: building a large model of online social networks with agents simulated by LLMs without the model being a TWON, i.e. a reliable copy of a network (c.f. Bail et al. (2023); Törnberg et al. (2023)).

However, it remains currently open if a model without being a twin of an OSN will allow generating knowledge about a particular OSN without being sufficiently approximate to the target system. Additionally, as we discussed in Section 5, it is uncertain if or not these governance modes are practically feasible and stable. If it turns out that all governance modes by which the main benefits of TWONs (protection of democratic institutions) can be achieved are practically infeasible, this provides an argument for the premise (4). This argument is depicted in Figure 7 as <A-GM-1-Pro-1-2: Impossibility of TWON's Regulation>. For this reason, this premise remains uncertain.

The Balancing Pro-Argument (<A-GM-1-Pro-2>), too, contains the premise (4) that there are no reasonable alternatives to GM-1. GM-2 (discussed in Section 8.2) is an example for an alternative governance mode with similar benefits but smaller moral drawbacks. However, it is uncertain whether it is practically feasible (see Section 5.2.2). We consider the premise (4) of the Balancing Pro-Argument as uncertain therefore.

Figure 8 depicts how important for the overall dialectic the claim is that governing modes restricting access to TWON are practically infeasible (this claim is conclusion of the argument <A-GM-1-Pro-1-2: Impossibility of TWON's Regulation>. If it is true, it provides justification for the two arguments supporting GM-1 (<A-GM-1-Pro-1> and <A-GM-1-Pro-2>). Additionally – that depicts Figure 8 – this claim justifies direct objections to all other governance modes discussed here. For the main normative claims recommending a certain governance mode (T-GM-2, T-GM-3, T-GM-4) imply that the governance modes are practically feasible ("ought implies can"). But if they turn out to be infeasible, negation of their recommendation follows.

Additionally, the <A-GM-1-Pro-2> argument contains a general normative claim (premise (3)) which might appear as highly implausible but which we have evaluated as uncertain. The principle claims that protecting democracy has such a high moral value that a means by which it can be attained should be used even if it involves the possibility of causing democratic institutions to collapse and damaging other moral values. As the last resort, it might be reasonable to justify a choice by such a risk-prone principle. However, it remains unclear whether the conditions of the last resort (no alternatives, democratic institutions are in realistic danger of collapse under business as usual etc.) hold.

GM-1 opposing arguments The Debunking Counter-Argument contains the claim that an unrestricted access to the TWON will not bring about significant benefits (premise (1)). This implies that all the identified benefits will not materialize, particularly that:

- it is unlikely that an unrestricted access to the TWON enhances social welfare in a significant amount (non-(B-1-1));

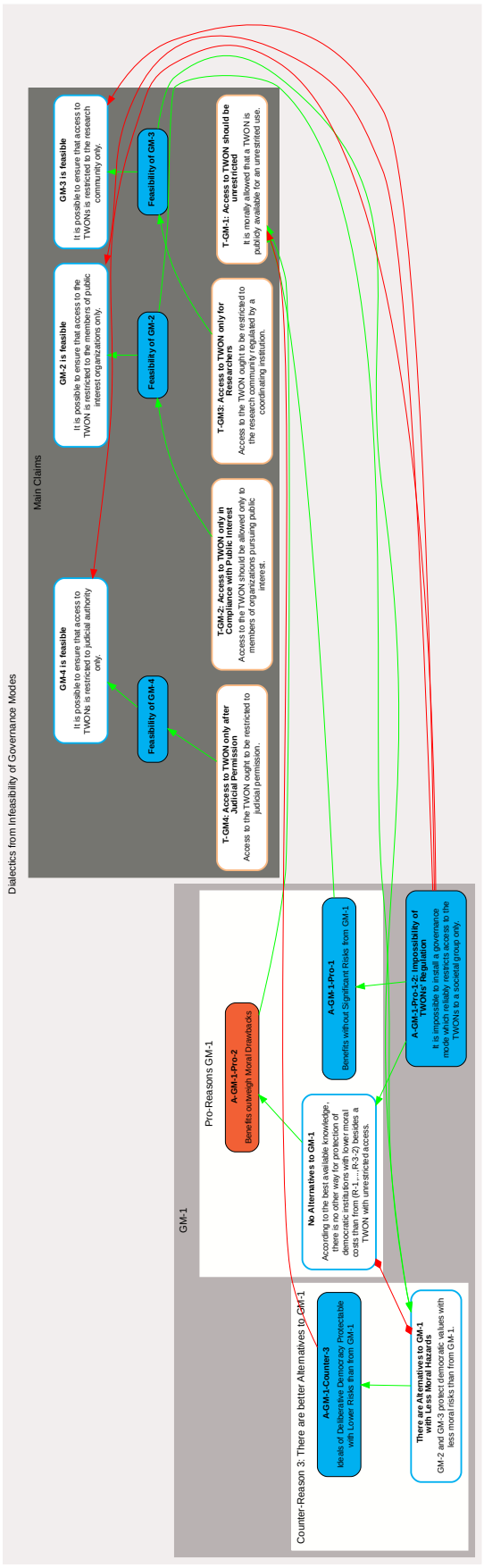

Figure 8: Dialectics from Infeasibility of Restricting Access to TWONs

- it is unlikely that it will improve individual quality of life (in an significant amount;
- it is unlikely that it will enable protection of democratic governance;
- it is unlikely that it will enable protection of individual autonomy on OSN;
- it is unlikely that it will significantly contribute to a mitigation of existing injustices.

We believe these statements to be highly implausible. We cannot imagine a reasonable idea how they could be justified.

The Moral-Thresholds Argument (<A-GM-1-Counter-2>) is based on a general principle which is a version of the so-called precautionary principles. In philosophical literature, precautionary principles are controversially discussed (e.g. Steel (2015)). For this reason, we have evaluated this premise as ”uncertain”.

The Alternatives-Argument (<A-GM-1-Counter-3>) is crucially based on the premise that there are ways to attain the most important benefits from the unrestricted access of TWON with smaller moral drawbacks (premise (4)). Possibly, models of OSN without being a twin of them will allow protection of democratic institutions and impose lower societal risks. Additionally, GM-2 and GM-3 do qualify as those alternatives according to our assessment (c.f. Section 8.2 for GM-2 and Section 8.3 for GM-3). However, as mentioned above, it is uncertain whether these governance modes are practically feasible and it is uncertain, if models of social media suffice. Accordingly, we consider this premise as uncertain.

Summary 6

A deductively valid reconstruction of arguments supporting and rejecting unrestricted access to the TWON reveals:

Pro-Arguments:

- Both pro-arguments contain uncertain premises. Additional inquiry is needed to evaluate their plausibility.
- Knowledge about practical feasibility of governance modes which restrict access to TWON is highly relevant for the argumentation. If they are infeasible, the two pro-arguments for GM-1 become much more plausible. If they are practically feasible, pro-arguments for GM-1 are rather implausible.
- Understanding of how severe current OSNs jeopardize democratic governance would provide knowledge necessary for evaluation of the first pro-argument for GM-1.

Counter-Arguments:

- The Debunking Counter-Argument rests on an implausible premise and is therefore implausible.
- The two further counter-arguments contain uncertain premises. Additional inquiry is needed to evaluate their plausibility.
- The Threshold Counter-Argument rests on a controversial normative decision principle (precautionary principle).
- Plausibility of the Alternatives-to-GM-1 counter-argument depends on whether or not the alternative governance modes are feasible.

8.2 GM-2: Public Interest Authority

GM-2 differs from GM-1 with respect to the number of societal members who will have access to the TWON. Whereas the GM-1 does not restrict access at all, the GM-2 allows accessing the TWON only to individuals on behalf of institutions which pursue public interest. Possible outcomes from GM-1 and GM-2 differ merely in degree: GM-2 prevents commercial use of TWON which will lead to lower economic benefits from TWON and lower likelihood that the risks from TWON – threads for individual autonomy, democratic governance, right to informational privacy, and perpetuation of existing injustices – will

materialize (since there is a smaller number of persons with access to the twin in the GM-2).

The qualitative differences in possible outcomes in GM-2 relative to GM-1 do not change the prima facie reasons supporting and objecting to this governance mode. Since all the benefits and all the risks remain and merely the probability of their occurrence changes in a non-quantifiable manner, the pro- and counter-reasons remain the same. Accordingly, the arguments supporting GM-2 and opposing it remain the same. Figure 9 depicts the dialectical structure as an argument map.

The claim [T-GM-2: Access to the TWON should be allowed only to members of organizations pursuing public interest.] can be justified by two arguments:

- Argument-GM-2-Pro-1 Debunking of Counter-Reasons
- Argument-GM-2-Pro-2 Balancing Benefits over Risks

Arguments rejecting [T-GM-2] are:

- Argument-GM-2-Counter-1: Debunking Benefits
- Argument-GM-2-Counter-2: Risks Transgress Moral Thresholds
- Argument-GM-2-Counter-3: Alternatives to GM-2

Evaluation of the arguments and the dialectical situation is the same as for GM-1. Pro-Arguments:

- The Debunking-Argument (<A-GM-2-Pro-1>) is committed to the claim that allowing access to public interest organizations will not lead to substantial moral drawbacks. In light of moral risks identified in Section 5, this claim requires additional support. As indicated in the previous section, additional inquiry is needed to assess plausibility of supporting arguments. At the current level of knowledge, we evaluate the premise (2) as ”uncertain”.
- The Balancing-Pro-Argument contains the general principle (means to protect democracy legitimize potential violation of other moral values) and the premise that there are no alternatives with lower moral threads. Both premises we consider as ”implausible”.

Counter-Arguments:

- <A-GM-2-Counter-1> contains a premise which denies that the TWON will realize substantial benefits. We consider that as ”implausible”.
- <A-GM-2-Counter-2>: it contains a version of the precautionary principle as the general principle. Since the precautionary principle is contested in the theoretical literature, we consider this premise as ”contested”.

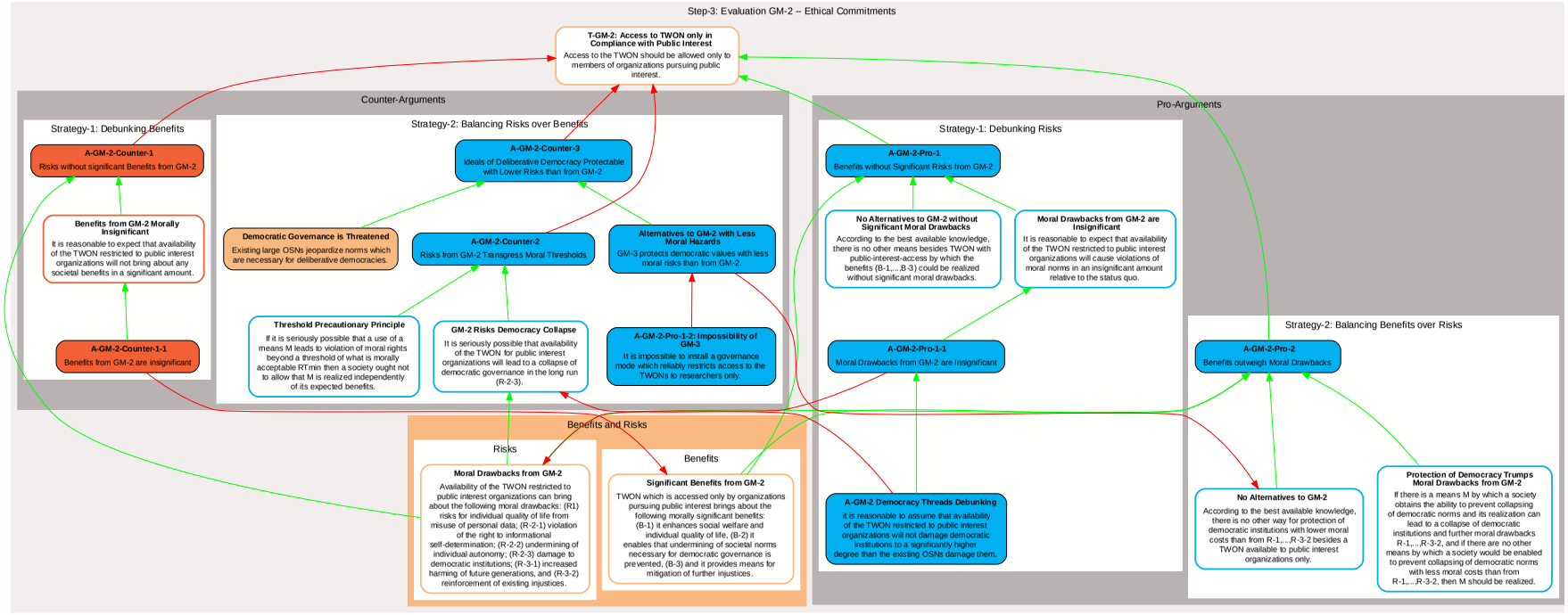

Figure 9: Dialectics about GM-2

Arguments and statements in orange represent starting assumptions of the argument analysis: the two statements (justified in Section 5.2) that a TWON according to GM-2 will bring about certain benefits and risks. At the top of the figure, the normative statement of interest [T-GM-2] is depicted by a white box with an orange line. Additionally, the argument <Democratic Governance is Threatened> which justifies a premise in the balancing-argument <A-GM-2-Counter-3> has been implicitly assumed at the outset of this study (c.f. Section 1 and references there for its elaboration).

The main statement [T-GM-2] is supported by two arguments depicted in the gray box ”Pro-Arguments” on the right and and attacked by three arguments depicted in the gray box ”Counter-Arguments” on the left.

Arguments and statements which we have evaluated as ”implausible” are coloured red; arguments and statements in blue we have evaluated as ”uncertain”. The two green arguments are considered as plausible.

- <A-GM-2-Counter-3>: this argument is based on the idea that there are alternative means by which the benefits from the respective governance mode will be reaped with lower drawbacks (c.f. p. 46). To support T-GM-2, this argument (depicted as <A-GM-2-Counter-3> in Figure 9) contains the premise that there are alternatives by which democratic governance can be protected without involving the moral drawbacks from GM-2. This claim, in turn, is justified by existence of more restrictive governance modes, e.g. GM-3, which protect democratic institutions as well but involve lower moral drawbacks than GM-2 (depicted as <Alternatives to GM-2 with less moral hazards> in Figure 9). However, if GM-3 is practically infeasible, this claim turns out to be false and a justification of <A-GM-2-Counter-3> fails. For this reason we have evaluated the argument <A-GM-2-Counter-3> as ”uncertain”.

  Infeasibility of a restricting access to TWONS plays, to the contrary, an opposing dialectical role for GM-1: Infeasibility of restrictive governance modes justifies premises in the two arguments supporting GM-1 (c.f. Figure 7).

A choice between GM-1 and GM-2 only cannot be justified at this stage. For that, a more precise description of possible outcomes and risks is needed, particularly more precise assessment of the degree of differences between the outcomes of the two governance modes.

Summary 7

A deductively valid reconstruction of arguments supporting and rejecting the governance mode according to which access to the TWON will be allowed only to members of organizations pursuing public interest reveals:

Pro-Arguments:

- The Balancing Pro-Argument rests on two implausible premises which makes the argument implausible.
- The Debunking Pro-Argument rests on two uncertain premises. Therefore, additional inquiry is needed to evaluate this argument's plausibility.

Counter-Arguments:

- The Debunking Counter-Argument rests on an implausible premise and is therefore implausible.
- The Threshold Counter-Argument rests on a controversial general principle (precautionary principle). Additionally, the argument contradicts the Debunking Pro-Argument whose plausibility is uncertain as well. The controversy revolves around the claim that it is seriously possible that availability of the TWON to public interest organizations will lead to a collapse of democratic institutions.
- The Alternatives-to-GM-2 Counter-Argument is uncertain.

8.3 GM-3: Research Community

This section presents reconstructed arguments directed at the claim [T1-GM-3: Access to the TWON ought to be restricted to the research community regulated by a coordinating institution.]. Premises which we consider to be "implausible" are highlighted in red. Premises which we consider to be "uncertain" are highlighted in violet. We justify our evaluations in Section 8.3.3.

8.3.1 Arguments Supporting GM-3

Argument-GM-3-Pro-1: Debunking of Counter-Reasons

(1) [Moral Benefits from GM-3] It is reasonable to expect that an access to TWON constrained to researchers under supervision of a coordinating institution (CI) will bring about the following morally significant benefits:

- (B-1-1) it enhances social welfare (in an amount of ΔSW(GM-3));

- (B-1-2) it improves individual quality of life (in an amount of ΔQoL(GM-3));
- (B-2-1) it enables that undermining of societal norms necessary for democratic governance is prevented;
- (B-3) it will contribute to a mitigation of existing injustices.

(2) [Moral Drawbacks from GM-3 are Insignificant]: It is reasonable to expect that access to the TWON constrained to researchers under supervision of a CI will lead to additional violations of moral rights in a morally insignificant amount.

(3) [Principle] A means M should be used if the following conditions are met:
(i) it is reasonable to expect that use of M brings about morally significant benefits B and
(ii) it is reasonable to expect that the moral drawbacks from M's use are insignificant and
(iii) for all other available means M_{alt} for which it is reasonable to expect that they realize at least B it is also reasonable to expect that they bring about moral drawbacks in a significant amount.

(4) [No-Alternatives to GM-3]: According to the best available knowledge, all other means by which the benefits (B-1-1,B-1-2,B-2-1, and B-3) could be realized involve significant moral drawbacks.

---

(5) [Conclusion]: T1-GM-3: Access to the TWON ought to be restricted to the research community regulated by a coordinating institution.

Argument-GM-3-Pro-2: Balancing Benefits over Risks

(1) [Moral Benefits from GM-3] It is reasonable to expect that an access to TWON constrained to researchers under supervision of a coordinating institution (CI) will bring about the following morally significant benefits:

- (B-1-1) it enhances social welfare (in an amount of ΔSW(GM-3));
- (B-1-2) it improves individual quality of life (in an amount of ΔQoL(GM-3));
- (B-2-1) it enables that undermining of societal norms necessary for democratic governance is prevented;
- (B-3) it will contribute to a mitigation of existing injustices.

(2) [Moral Drawbacks from GM-3]: It is reasonable to expect that an access to TWON restricted to researchers under supervision of CI will involve moral drawbacks in a morally significant amount:

- (R-2-1) Reinforcement of violations of the right to informational privacy;
- (R-3) Reinforcement of existing injustices.

(3) [Protection of Democracy Trumps Moral Drawbacks]: Protection of democracy has such a high moral importance that it holds:
if there is a means M for which it is reasonable to expect that it provides the ability to prevent collapsing of democratic norms but which reinforces violations of the right to informational privacy (R-2-1) and it reinforces existing injustices (R-3)
and if for all other available means M_{alt} for which it is reasonable to expect that they protect democratic institutions it is reasonable to expect that they will involve more severe moral drawbacks than (R-2-1 and R-3),
then M ought to be realized.

(4) [No Alternatives to GM-3]: According to the best available knowledge, all alternative ways to protect democratic institutions involve higher moral drawbacks than R-2-1 and R-3.

---

(5) [Conclusion]: T1-GM-3: Access to the TWON ought to be restricted to the research community regulated by a coordinating institution.

8.3.2 Arguments Objecting to GM-3

Argument-GM-3-Counter-1: Debunking-Benefits

(1) [Benefits from GM-3 Morally Insignificant]: It is reasonable to expect that an access to the TWON restricted to researchers will not bring about any societal benefits in a significant amount.

(2) [Moral Drawbacks from GM-3]: It is reasonable to expect that an access to TWON constrained to researchers under supervision of CI will involve moral drawbacks in a morally significant amount:

- (R-2-1) Reinforcement of violations of the right to informational privacy;
- (R-3) Reinforcement of existing injustices.

(3) [Principle]: If it is reasonable to expect that a means M jeopardizes moral rights to a morally significant degree and it is reasonable to expect that M will not realize morally significant benefits, it is morally not allowed to realize M.

---

(4) Conclusion: Non-T1-GM-3: Research community regulated by a coordinating institutions ought not to have access to the TWON.

Argument-GM-3-Counter-2: Risks from GM-3 Transgress Moral Thresholds

(1) [Threshold Precautionary Principle]: If it is seriously possible that a use of a means M leads to violation of moral rights beyond a threshold of what is morally acceptable R_{min} then a society ought not to allow that M is realized independently of its expected benefits.

(2) [Moral Drawbacks from GM-3]: It is seriously possible that an access to TWONs restricted to researchers under supervision of CI will involve moral drawbacks in a morally significant amount:

- (R-2-1) Reinforcement of violations of the right to informational privacy;
- (R-3) Reinforcement of existing injustices.

(3) Reinforcement of existing injustices and of additional violations of the right to informational privacy are outcomes beyond a threshold of what is morally acceptable.

---

(4) Conclusion: Non-T1-GM-3: Research community regulated by a coordinating institutions ought not to have access to the TWON.

Argument-GM-3-Counter-3: Alternatives to GM-3

(1) There is a prima facie moral requirement that a democratic society has the ability to protect societal norms necessary for democratic governance.

(2) From all available means $M_1, ..., M_n$ for which it is reasonable to expect that they protect democratic norms, only the means M_i ought to be chosen for which it is reasonable to expect that its realization has the lowest moral drawbacks.

(3) [Benefits and Drawbacks from GM-3]: It is reasonable to expect that access to TWON restricted to researchers will enable protection of democratic institutions (B-2-1) and realize further benefits (B-1-1, B-1-2,B-3) and that it involves moral drawbacks (R-2-1 and R-3).

(4) [Alternatives for Democracy Protection with Lower Moral Drawbacks Available]: There are means for which it is reasonable to expect that they protect democratic norms and reinforce existing injustices and violations of the right to informational privacy to a lesser degree than GM-3-TWON:

<+ It is reasonable to expect that analysis of OSN-data obtained via the Digital Services Act will enable protection of democratic norms without violation of privacy or amplifying existing injustices.

(5) [Preliminary Conclusion from (3)&(4)]: It is not the case that it is reasonable to expect that access to the TWON restricted to researchers protects democracy with the lowest moral drawbacks.

(6) [Conclusion from 1&2&5]: Non-T1-GM-3: Research community regulated by a coordinating institutions ought not to have access to the TWON.

Figure 10 depicts the resulting dialectical structure.

8.3.3 Discussion

GM-3 supporting arguments The Debunking-Pro-Argument (<A-GM-3-Pro-1>) contains the claim that TWONs whose access is restricted to researchers according to the rules of the governance mode 3 do not involve significant moral risks (premise (2)). This implies that all the risks identified in Section 5.3 for GM-3 will most likely not materialize. Thereby, a proponent of the Debunking-Pro-Argument is committed to the following claims:

- It is unlikely that access to the TWON restricted to researchers will violate the right to informational self-determination or the right to informational self-determination does not hold;
- it is unlikely that an access to the TWON restricted to researchers will reinforce existing injustices.

These claims demand a justification (depicted by the argument <A-GM-3-Pro-1-1> in Figure 10). Here are some ideas how these claims could be justified:

- to account against violations of the right to informational privacy, the coordinating institution (CI) of the TWON could demand that research projects are conducted on anonymized data. Alternatively, it could be the case that the existing OSNs are already violating the right to informational privacy to such degree that a researcher's access to a TWON will not significantly deteriorate the situation.
- We can foresee only one injustice which can be fortified by a TWON restricted to researchers: global epistemic inequalities. This challenge can, in principle, be accommodated. There are different ways how it could be secured that researchers from all parts of the world obtain access to the TWON if they meet the access conditions.

These ideas need to be substantiated for which additional research is needed. Currently, we consider the premise (2) of the Debunking-Pro-Argument as uncertain.

The Balancing Pro-Argument (<A-GM-3-Pro-2>) contains two normative premises which we evaluate as "uncertain":

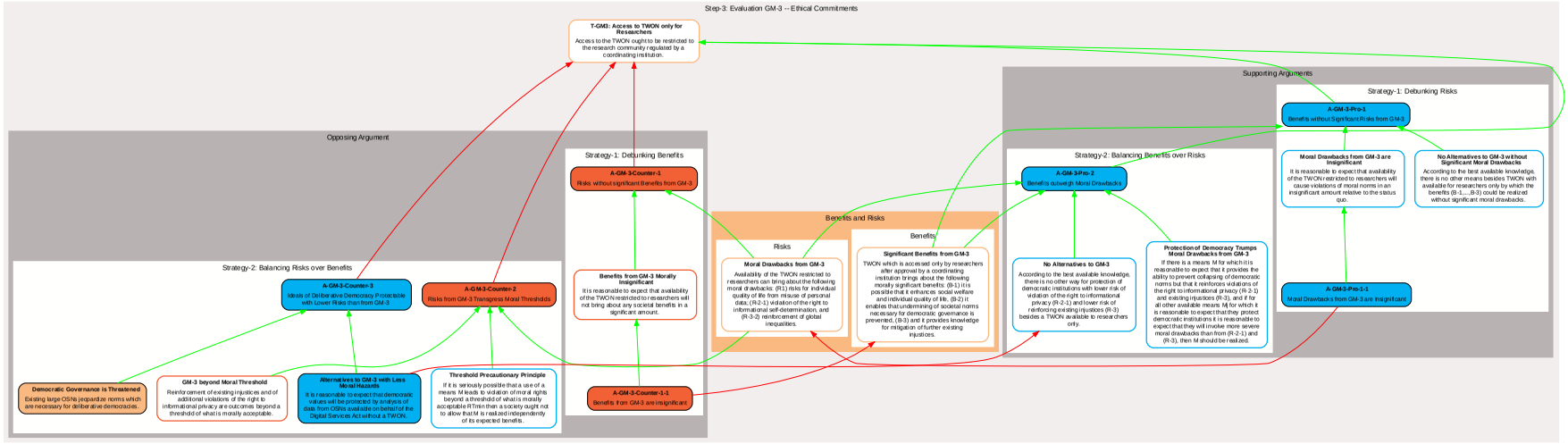

Figure 10: Dialectics about GM-3

Arguments and statements in orange represent starting assumptions of the argument analysis: the two statements (justified in Section 5.3) that a TWON according to GM-3 will bring about certain benefits and risks (orange box in the middle). At the top of the figure, the normative statement of interest [T-GM-3] is depicted by a white box with an orange line. Additionally, the argument <Democratic Governance is Threatened> (on the left) which justifies a premise in the balancing-argument <A-GM-3-Counter-3> has been implicitly assumed at the outset of this study (c.f. Section 1 and references there for its elaboration).
The main statement [T-GM-3] is supported by two arguments depicted in the gray box ”Pro-Arguments” on the right and and attacked by three arguments depicted in the gray box ”Counter-Arguments” on the left.
Arguments and statements which we have evaluated as ”implausible” are coloured red; arguments and statements in blue we have evaluated as ”uncertain”.

- The general principle in premise (3): It claims that protection of democratic institutions has such moral importance that violations of the right to informational privacy and reinforcement of further injustices ought to be tolerated. Plausibility of this claim requires a more precise quantification of the involved risks to moral rights. At the current stage of knowledge, its plausibility remains undecided.

- Premise (4) of the Debunking-Pro-Argument claims the that there are no appropriate alternatives to GM-3. As discussed on p. 64, large models of social media might turn out as an appropriate alternative to a TWON. However, we cannot decide this on the current state of knowledge. For this reason, we evaluate this premise as ”uncertain” as well. Its opposite, the claim that there are appropriate alternatives, is contained in the Argument-GM-3-Counter-3 (premise (4)) with a proposal for an alternative. We consider this controversy as undecided at the current state of research.

GM-3 opposing arguments The Debunking Counter-Argument contains the claim that an access to the TWON restricted to researchers will not bring about significant benefits (premise (1)). This implies that all the identified benefits (Section 5.3 will not materialize. We do not see how these claims can be reasonably justified. For this reason we have evaluated the premise (1) as ”implausible”.

The Moral-Thresholds Argument (<A-GM-3-Counter-2>) is based on a general principle which is a version of precautionary principles which are controversially discussed in philosophical literature (e.g. Steel (2015)). For this reason, we have evaluated this premise as ”uncertain”. However, the Moral-Thresholds-Argument contains a premise which we consider as ”implausible”. According to it, the risks from GM-3 – violations of the right to informational privacy and reinforcement of existing injustices – transgress the threshold of what is morally acceptable.

The Alternatives-Argument (<A-GM-3-Counter-3>) is crucially based on the premise that there are ways to attain the most important benefits from the unrestricted access of TWON with smaller moral drawbacks (premise (4)). Evaluation of this claim requires additional inquiry analysing the question of whether simpler models than a TWON would allow reaching the same normative goals with lesser moral drawbacks.

Summary 8

A deductively valid reconstruction of arguments supporting and rejecting only-researchers access to TWONs reveals:

Pro-Arguments: Both supporting arguments – Debunking-Risks and Balancing-Benefits-over-Risks – rest on uncertain premises. Additional inquiry is needed to evaluate plausibility of both arguments.

Counter-Arguments:

- The Debunking Counter-Argument rests on an implausible premise and is therefore implausible.
- The Threshold Counter-Argument rests on a controversial general principle (precautionary principle) and an implausible premise. The argument is therefore implausible.
- The Alternatives-to-GM-2 Counter-Argument rests on an uncertain premise. Additional inquiry is needed to evaluate it.

8.4 GM-4: Judicial Authority

GM-4 is the most restrictive governance mode of the discussed ones. It provides access to the TWON only on behalf of a legal authority for investigation of certain illegal acts or outcomes. Accordingly, the benefits and the risks from this governance mode are lower than the benefits and the risks from the other governance modes (c.f. Section 5.4). This section presents the reconstructed arguments directed at the claim [T-GM-4: Access to the TWON ought to be restricted to agents with a judicial permission.] and its negation.

8.4.1 Arguments Supporting GM-4

Argument-GM-4-Pro-1: Debunking of Counter-Reasons

(1) [Moral Benefits from GM-4] It is reasonable to expect that an access to TWON constrained to judicial permission enables a society to prevent and prosecute violations of legal norms on OSNs.

(2) [Moral Drawbacks from GM-4 are Insignificant]: It is reasonable to expect that access to the TWON constrained to juridical directive will not lead to violations of moral rights in a morally significant amount relative to the status quo.

(3) [Principle] A means M should be used if the following conditions are met:
(i) it is reasonable to expect that use of M brings about morally significant benefits B and
(ii) it is reasonable to expect that the moral drawbacks from M's use are insignificant and
(iii) for all other available means M_{alt} for which it is reasonable to expect that they realize at least B it is also reasonable to expect that they bring about moral drawbacks in a significant amount.

(4) [No-Alternatives to GM-4]: According to the best available knowledge, there is no other means by which legal norms on OSNs will be enforced without significant moral drawbacks.

---

(5) [Conclusion]: T-GM-4: Access to the TWON ought to be restricted to agents with a judicial permission.

Argument-GM-4-Pro-2: Balancing Benefits over Risks

(1) [Enforcement of Legal Norms by TWON-GM-4] It is reasonable to expect that TWON which is accessed only on behalf of a judicial authority enables a society to prevent and prosecute violations of legal norms on OSNs.

(2) [Violation of the Privacy Right by TWON-GM-4]: It is reasonable to expect that an access to TWON restricted to judicial directive will reinforce violations of the right to informational privacy.

(3) [Law Enforcement Trumps Privacy]: Ability to enforce law on OSNs has such a high moral importance that it holds:
If there is a means M by which a society obtains the ability to enforce law on OSNs but which violates the right to informational self-determination
and for all other available means M_{alt} for which it is reasonable to expect that they enforce law on OSNs it is reasonable to expect that they will involve more severe moral drawbacks than violations of the right to informational self-determination,
then M ought to be realized.

(4) [No Alternatives to GM-4]: According to the best available knowledge, all alternative ways to enforce law on OSNs involve higher moral drawbacks than violations of the right to informational privacy.

---

(5) [Conclusion]: T-GM-4: Access to the TWON ought to be restricted to agents with a judicial permission.

8.4.2 Arguments Objecting to GM-4

Argument-GM-4-Counter-1: Debunking-Benefits

(1) [TWON-GM-4 will not Enforce Laws]: It is reasonable to expect that a TWON accessible on behalf of a judicial authority only will not enforce legal norms on OSNs.

(2) [Violation of the Privacy Right by TWON-GM-4]: It is reasonable to expect that an access to TWON restricted to judicial directive will reinforce violations of the right to informational privacy.

(3) [Principle]: If it is reasonable to expect that a means M jeopardizes moral rights to a morally significant degree and it is reasonable to expect that M will not realize morally significant benefits, it is morally not allowed to realize M.

---

(4) Conclusion: non-T-GM-4: It is morally not allowed that access to the TWON is restricted to agents with a judicial permission.

Argument-GM-4-Counter-2: Risks Transgress Moral Thresholds

(1) [Threshold Precautionary Principle]: If it is seriously possible that a use of a means M leads to violation of moral rights beyond a threshold of what is morally acceptable R_{min} then a society ought not to allow that M is realized independently of its expected benefits.

(2) [Violations of Privacy Rights Seriously Possible]: It is seriously possible that an access to TWON restricted to judicial directive will reinforce violations of the right to informational privacy.

(3) Reinforcement of violations of the right to informational privacy is an outcome beyond the threshold of what is morally acceptable.

---

(4) Conclusion: non-T-GM-4: It is morally not allowed that access to the TWON is restricted to agents with a judicial permission.

Argument-GM-4-Counter-3: Alternatives to GM-4

(1) There is a prima facie moral requirement that illegal acts on OSNs ought to be prosecuted by law.

(2) From all available means $M_1, ..., M_n$ for which it is reasonable to expect that they enable law enforcement on OSNs, only the means M_i ought to be chosen for which it is reasonable to expect that its realization has the lowest moral drawbacks.

(3) [Benefits and Drawbacks from GM-4]: It is reasonable to expect that access to TWON restricted to judicial directives will enable law enforcement on OSNs and it is reasonable to expect that this governance mode will reinforce violations of the right to informational privacy.

(4) [Alternatives for Law Enforcement with Lower Moral Drawbacks Available]: There are means for which it is reasonable to expect that they enable law enforcement on OSNs and violate the right to informational privacy to a lesser degree than GM-4-TWON:

<+ It is reasonable to expect that research with toy-models based on OSN-data provided via Digital Services Act will enable prosecution of illegal acts on OSNs without violation of privacy.

---

(5) [Preliminary Conclusion from (3)&(4)]: It is not the case that it is reasonable to expect that access to the TWON restricted to judicial directives enables laws enforcement on OSNs with the lowest moral drawbacks.

---

(6) [Conclusion from 1&2&5]: Non-T-GM-4: It is morally not allowed that access to the TWON is restricted to agents with a judicial permission.

Figure 11 depicts the resulting dialectical structure.

8.4.3 Discussion

As the argumentative map in Figure 11 depicts, the recommendation of GM-4 is supported by two arguments which are currently uncertain and it is attacked by one argument which rests on uncertain premises. The other two counter-arguments are implausible already in light of the available knowledge about its risks and benefits.

The main drawback of this governance mode is, however, the fact that GM-4 does not attain one crucial goal: protection of democratic institutions. Thus, even if GM-4 turns out to be well justified, the

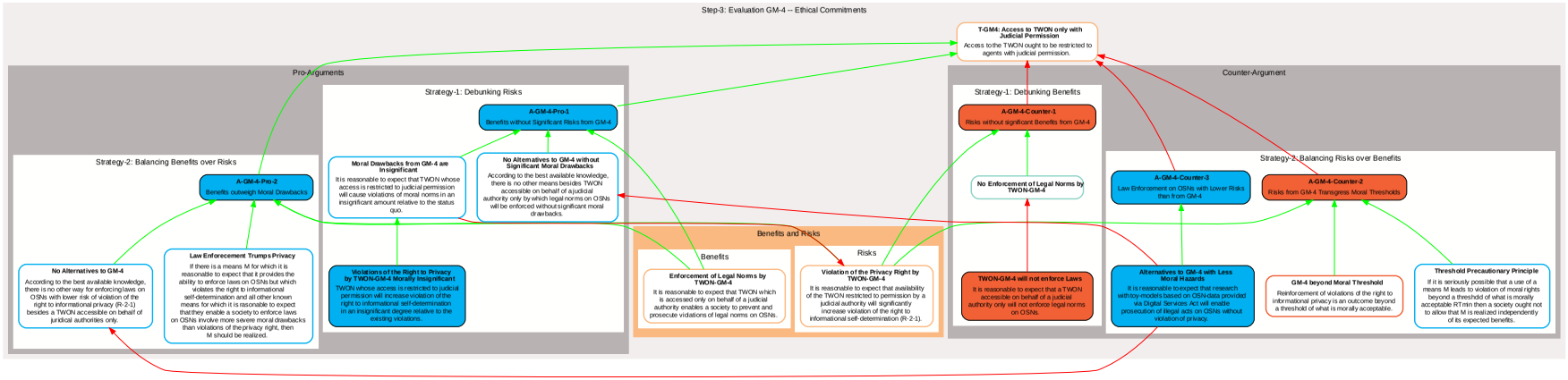

Figure 11: Dialectics about GM-4

Arguments and statements in orange represent starting assumptions of the argument analysis: the two statements (justified in Section 5.3) that a TWON according to GM-4 will bring about certain benefits and risks (orange box in the middle). At the top of the figure, the normative statement of interest [T-GM-4] is depicted by a white box with an orange line. The main statement [T-GM-4] is supported by two arguments depicted in the gray box ”Pro- Arguments” on the right and and attacked by three arguments depicted in the gray box ”Counter-Arguments” on the left. Arguments and statements which we have evaluated as ”implausible” are coloured (arguments) or framed (statements) red; arguments and statements in blue we have evaluated as ”uncertain”.

question will remain whether a less restrictive access to TWON is legitimate in order to protect democratic values.

GM-4 Supporting Arguments Both arguments rest on premises which are currently uncertain.

The Debunking-Pro-Argument (<A-GM-4-Counter-1>) presupposes the claim (premise (2)) that GM-4 will not violate the right to informational privacy to a substantial degree. This claim needs to be evaluated when it becomes clear where TWON's data come from. The Balancing-Pro-Argument (<A-GM-4-Pro-2>) contains a balancing principle (premise (3)) and the claim about none-existence of suitable alternatives (premise (4)). Both claims require more detailed inquiries to evaluate their truth-value (or at least plausibility).

GM-4 Opposing Arguments The Debunking-Counter-Argument (<A-GM-4-Counter-1>) is implausible since it contains the premise (1) which claims that the expected outcome of the GM-4 – persecution of illegal acts – is morally insignificant. It is unclear how this claim could be reasonably justified.

The Moral-Thresholds-Argument (<A-GM-4-Counter-2>) is implausible because premise (3) claims that violations of the right to informational self-determination due to TWON development and use transgress the threshold of what is morally acceptable.

The Alternatives-Argument (<A-GM-4-Counter-3>) relies on the claim that acceptable alternatives to GM-4 exist. As discussed above, we currently do not know such alternatives. But we cannot exclude that with original data from OSNs, legal prosecution will be possible without a digital twin of the OSN. Or, it might turn out that large models of online social media – "Social Media Accelerator" (Bail et al., 2023) – allows enforcing laws. For this reason, we evaluate the argument as "uncertain" to point out that it needs further inquiries.

Summary 9

A deductively valid reconstruction of arguments supporting and rejecting access to TWONs restricted to juridical permission (GM-4) reveals:

- This governance mode does not attain a main target of TWON's development: identification of design choices in social network's platforms which influence communicative dynamics towards outcomes which violate ethical norms necessary for democratic governance. Instead, it pursues the goal of enforcing laws on OSNs.
- Both arguments supporting this governance mode are controversial. Additional inquiry is needed to evaluate them.
- Two arguments rejecting this governance mode are implausible (Debunking-Benefits and Threshold-Transgression). GM-4 can be rejected if there is an alternative way to enforce laws on OSNs with lower moral drawbacks than TWON. Currently, it remains uncertain if such an alternative exists.

References

Christopher A Bail, D Sunshine Hillygus, Alexander Volfovsky, Max Allamong, Fatima Alqabandi, Diana ME Jordan, Graham Tierney, Christina Tucker, Andrew Trexler, and Austin van Loon. Do we need a social media accelerator? *SocArXiv doi*, 10, 2023. URL `https://files.osf.io/v1/resources/ucfbk/providers/osfstorage/6585bb7a7094e91fa2a17407`. 64, 82

Balbir S. Barn. The sociotechnical digital twin: On the gap between social and technical feasibility. In *2022 IEEE 24th Conference on Business Informatics (CBI)*. IEEE, June 2022. doi: 10.1109/cbi54897.2022.00009. 9

Peter Bauer, Bjorn Stevens, and Wilco Hazeleger. A digital twin of earth for the green transition. *Nature Climate Change*, 11(2):80–83, 2021. ISSN 1758-6798. doi: 10.1038/s41558-021-00986-y. 10

Gregor Betz. *Argumentationsanalyse. Eine Einführung*. J.B. Metzler, Berlin, 2020. 15

Gregor Betz. Natural-language multi-agent simulations of argumentative opinion dynamics. *Journal of Artificial Societies and Social Simulation*, 25(1), 2022. ISSN 1460-7425. doi: 10.18564/jasss.4725. 9

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S

Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. 28

Georg Brun and Gregor Betz. Analysing practical argumentation. In Sven Ove Hansson and Gertrude Hirsch Hadorn, editors, *The Argumentative Turn in Policy Analysis*, chapter 3, pages 39–78. Springer International Publishing, 2016. doi: 10.1007/978-3-319-30549-3_3. URL `http://dx.doi.org/10.1007/978-3-319-30549-3_3`. 15, 20

Axel Bruns. *Are filter bubbles real?* John Wiley & Sons, 2019. 9

European Commission. Living guidelines on the responsible use of generative AI in research, 2024. URL `https://research-and-innovation.ec.europa.eu/document/download/2b6cf7e5-36ac-41cb-aab5-0d32050143dc_en?filename=ec_rtd_ai-guidelines.pdf`. 14

Casey Fiesler and Nicholas Proferes. "participant" perceptions of twitter research ethics. *Social Media + Society*, 4(1):205630511876336, 2018. ISSN 2056-3051. doi: 10.1177/2056305118763366. 36

Andreas Flache, Michael Mäs, Thomas Feliciani, Edmund Chattoe-Brown, Guillaume Deffuant, Sylvie Huet, and Jan Lorenz. Models of social influence: Towards the next frontiers. *Journal of Artificial Societies and Social Simulation*, 20(4), 2017. 9, 12

A. Michael Froomkin. Big data: Destroyer of informed consent. *Yale Journal of Health Policy, Law, and Ethics*, 18(3):27–54, 2019. 36

Nigel Gilbert, Petra Ahrweiler, Pete Barbrook-Johnson, Kavin Preethi Narasimhan, and Helen Wilkinson. Computational modelling of public policy: Reflections on practice. *Journal of Artificial Societies and Social Simulation*, 21(1), 2018. ISSN 1460-7425. doi: 10.18564/jasss.3669. 27

Armin Grunwald. *Technology Assessment in Practice and Theory*. Routledge, Abingdon, 2018. ISBN 9780429442643. doi: 10.4324/9780429442643. 14

Thilo Hagendorff. The ethics of ai ethics: An evaluation of guidelines. *Minds and Machines*, 30(1):99–120, February 2020. ISSN 1572-8641. doi: 10.1007/s11023-020-09517-8. 14

Sven Ove Hansson. Risk. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2023 edition, 2023. 13

Maralee Harrell. *What Is the Argument? An Introduction to Philosophical Argument and Analysis*. MIT Press, Cambridge, Massachusetts, 2016. ISBN 9780262529273. 15

House of Representatives. Disinformation nation: Social media's role in promoting extremism and misinformation, 2021. URL https://www.congress.gov/event/117th-congress/house-event/111407. 8

Marijn A Keijzer and Michael Mäs. The complex link between filter bubbles and opinion polarization. *Data Science*, 5(2):139–166, 2022. 8, 9

Dominik Klein, Johannes Marx, and Kai Fischbach. Agent-based modeling in social science, history, and philosophy. an introduction. *Historical Social Research / Historische Sozialforschung*, 43(1):7–27, 2018. doi: 10.12759/HSR.43.2018.1.7-27. 24

David Lazer, Eszter Hargittai, Deen Freelon, Sandra Gonzalez-Bailon, Kevin Munger, Katherine Ognyanova, and Jason Radford. Meaningful measures of human society in the twenty-first century. *Nature*, 595(7866):189–196, June 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03660-7. 29

Helen Margetts and Cosmina Dorobantu. Computational social science for public policy. In Eleonora Bertoni, Matteo Fontana, Lorenzo Gabrielli, Serena Signorelli, and Michele Vespe, editors, *Handbook of Computational Social Science for Policy*, chapter 1, pages 3–18. Springer International Publishing, 2023. ISBN 9783031166242. doi: 10.1007/978-3-031-16624-2_1. 27, 29

Alison McIntyre. Doctrine of Double Effect. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2023 edition, 2023. 32

Stefan Mihai, Mahnoor Yaqoob, Dang V. Hung, William Davis, Praveer Towakel, Mohsin Raza, Mehmet Karamanoglu, Balbir Barn, Dattaprasad Shetve, Raja V. Prasad, Hrishikesh Venkataraman, Ramona Trestian, and Huan X. Nguyen. Digital twins: A survey on enabling technologies, challenges, trends and future prospects. *IEEE Communications Surveys (&) Tutorials*, 24(4):2255–2291, 2022. ISSN 2373-745X. doi: 10.1109/comst.2022.3208773. 27

Brent Daniel Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2):205395171667967, December 2016. doi: 10.1177/2053951716679679. 14, 24, 36

Adam D. Moore. Privacy: Its meaning and value. *American Philosophical Quarterly*, 40(3):215–227, 2003. ISSN 00030481. URL http://www.jstor.org/stable/20010117. 33

Barack Obama. Farewell adress, 2017. URL https://obamawhitehouse.archives.gov/farewell. 8

Eli Pariser. *The filter bubble: What the Internet is hiding from you*. Penguin UK, 2011. 8

Natahniel Persily and Joshua A. Tucker, editors. *Social Media and Democracy: The State of the Field, Prospects for Reform*. Cambridge University Press, Cambridge, UK, August 2020. ISBN 9781108812894. doi: 10.1017/9781108890960. 9, 32

Adil Rasheed, Omer San, and Trond Kvamsdal. Digital twin: Values, challenges and enablers from a modeling perspective. *IEEE access*, 8:21980–22012, 2020. 9

Beate Roessler. X—privacy as a human right. *Proceedings of the Aristotelian Society*, 117(2):187–206, July 2017. ISSN 1467-9264. doi: 10.1093/arisoc/aox008. 33

Thomas M. Scanlon. *What We Owe to Each Other*. The Belknap Press of Harvard University Press, Cambridge, Massachusetts, and London, England, 1998. 22

Pawel Sobkowicz. Social simulation models at the ethical crossroads. *Science and Engineering Ethics*, 25(1):143–157, November 2017. doi: 10.1007/s11948-017-9993-0. 27

Daniel Steel. *Philosophy and the Precautionary Principle*. Cambridge University Press, Cambridge, 2015. ISBN 1107078164. URL `http://www.ebook.de/de/product/22587528/daniel_steel_philosophy_and_the_precautionary_principle.html`. 67, 77

Franz-Walter Steinmeier. Christmas message, 2017. URL `https://www.bundespraesident.de/SharedDocs/Reden/EN/FrankWalter-Steinmeier/Reden/2018/12/181225-Christmas-message.html`. 8

Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in NLP. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1355. URL `https://aclanthology.org/P19-1355`. 28

Daniel Susser, Beate Roessler, and Helen Nissenbaum. Technology, autonomy, and manipulation. *Internet Policy Review*, 8(2), June 2019. ISSN 2197-6775. doi: 10.14763/2019.2.1410. 32

Holm Tetens. *Philosophisches Argumentieren*. Beck'sche Reihe. C.H. Beck, München, 2004. 15

Petter Törnberg, Diliara Valeeva, Justus Uitermark, and Christopher Bail. Simulating social media using large language models to evaluate alternative news feed algorithms. 2023. doi: 10.48550/arXiv.2310.05984. URL `https://doi.org/10.48550/arXiv.2310.05984`. 64

Jeroen van den Hoven, Martijn Blaauw, Wolter Pieters, and Martijn Warnier. Privacy and Information Technology. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2020 edition, 2020. 33, 36

Louise Wright and Stuart Davidson. How to tell the difference between a model and a digital twin. *Advanced Modeling and Simulation in Engineering Sciences*, 7(13), March 2020. ISSN 2213-7467. doi: 10.1186/s40323-020-00147-4. 9

Carole-Jean Wu, Ramya Raghavendra, Udit Gupta, Bilge Acun, Newsha Ardalani, Kiwan Maeng, Gloria Chang, Fiona Aga, Jinshi Huang, Charles Bai, Michael Gschwind, Anurag Gupta, Myle Ott, Anastasia Melnikov, Salvatore Candido, David Brooks, Geeta Chauhan, Benjamin Lee, Hsien-Hsin Lee, Bugra Akyildiz, Maximilian Balandat, Joe Spisak, Ravi Jain, Mike Rabbat, and Kim Hazelwood. Sustainable ai: Environmental implications, challenges and opportunities. In D. Marculescu, Y. Chi, and C. Wu, editors, *Proceedings of Machine Learning and Systems*, volume 4, pages 795–813, 2022. URL `https://proceedings.mlsys.org/paper_files/paper/2022/file/462211f67c7d858f663355eff93b745e-Paper.pdf`. 28

Ekaterina Zhuravskaya, Maria Petrova, and Ruben Enikolopov. Political effects of the internet and social media. *Annual review of economics*, 12:415–438, 2020. 9

Contact us:

Damian Trilling
Project Coordinator

+31 62 782 7904

d.c.trilling@uva.nl

University of Amsterdam
Postbus 15791
1001 NG Amsterdam