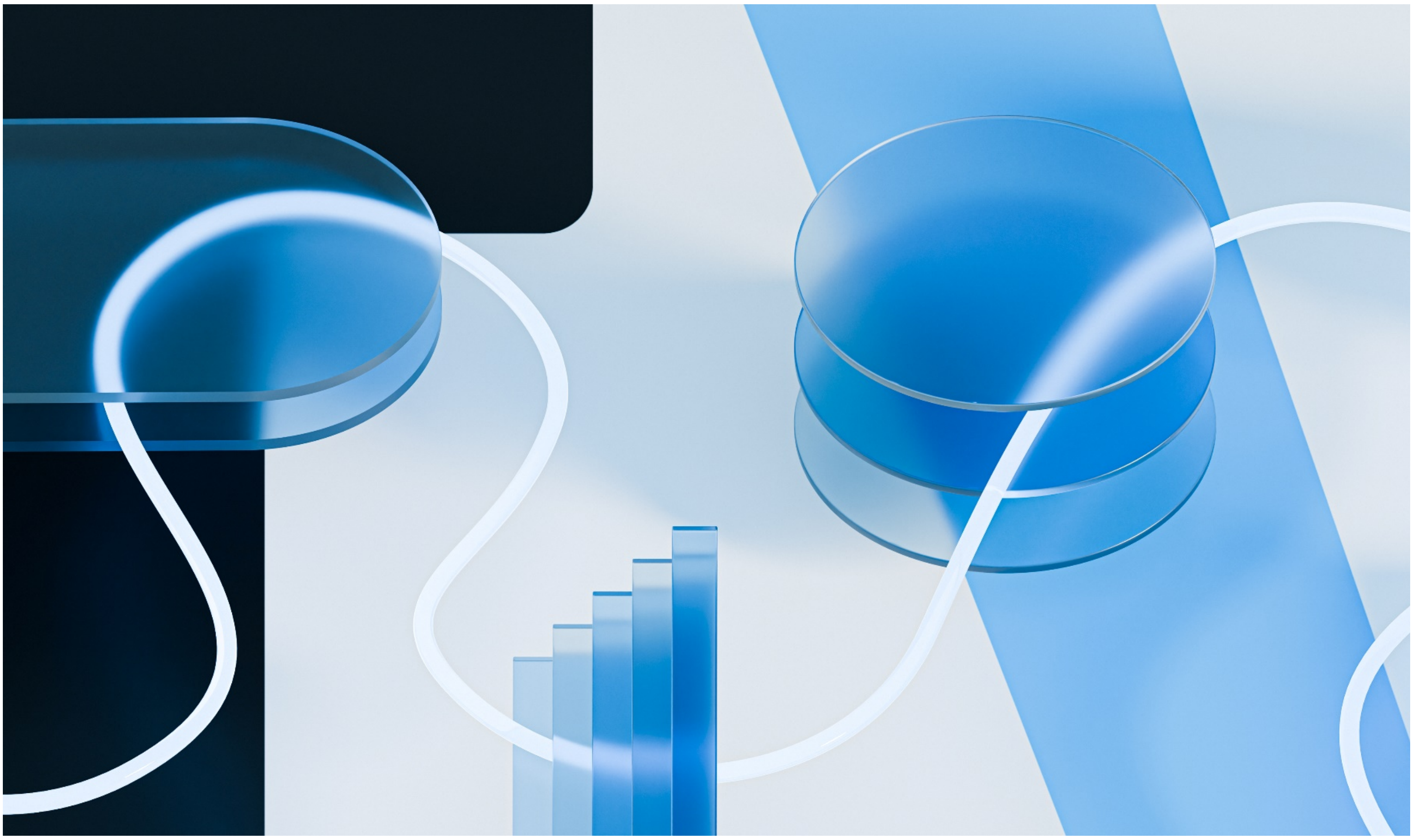

TWin of Online Social Networks

Deliverable D2.1

General TWON Prototype

Main Authors: Michael Mäs, Fabio Sartori, Andreas Reitenbach, Alenka Gucek, Abdul Sittar, Simon Münker

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), Dialogue Perspectives e.V (Germany).

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	D2.1
Deliverable Name	General TWON prototype
Due Date	31.03.2025
Submission Date	31.03.2025
Type	DEM — Demonstrator, pilot, prototype
Dissemination Level	PU - Public
Work Package	WP 2
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), Dialogue Perspectives e.V (DIA)

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 the present document, the TWON consortium is describing the first prototype of their TWON. This includes a description of our methodological approach, as well as the current version's assumptions and their technical implications.

Contents

List of Abbreviations 5

1 Introduction 6

1.1 Background 6

1.2 TWONs - Digital Twins of online social networks 7

1.3 TWON's Potentials 8

1.4 Ethical Background 10

1.5 Quality criteria for TWONs: realism and flexibility 10

2 How to develop a TWON? 11

2.1 Methodology 11

2.2 Example: Does success-driven user activity matter? 13

3 Overview over the TWON 17

3.1 Conceptual model 17

3.2 Computational model 19

4 The user model 22

4.1 Theoretical foundation 22

4.2 Resource budget 25

4.3 User activation 25

4.3.1 The decision to log on 25

4.3.2 Managing expectations 27

4.3.3 Time cost function 28

4.3.4 Social feedback reward function 29

4.3.5 Other factors 30

4.3.6 pop-up notifications 31

4.4 User platform actions, and deactivation (owner: Fabio) 31

4.4.1 New content generation (DONE) 33

4.4.2 Content consumption (OWNER: FABIO) 34

4.4.3 Interest and agreement 36

4.4.4 Personal value update 36

4.4.5 Liking/disliking a post or a comment 37

4.4.6 Forwarding/sharing a post or a comment 37

4.4.7 Commenting a post or a comment 38

4.4.8 Waiting 38

4.5 Content creation with LLMs 38

5 The platform model **41**

5.1 Ranking of incoming content 42

6 Outcome measures **44**

7 Technical infrastructure **45**

7.1 System overview 45

7.2 Simulation services 47

7.2.1 Network generators 48

7.2.2 Ranking 48

7.2.3 Agents Scheduler 49

7.3 Infrastructure and Deployment 49

7.3.1 Medium scale simulations 49

References **52**

List of Abbreviations

LLM	Large Language Model
OSN	Online Social Network
TWON	Twin of an Online Social Network

TWON prototype

Michael Mäs, Fabio Sartori, Andreas Reitenbach, Alenka Guček, Abdul Sittar, Simon Münker *

March 31, 2025

1 Introduction

The aim of the present document is to describe the TWON prototype, documenting the main modeling decisions, the modeling work, and the technical implementation of the past months' work of the TWON consortium. In the following subsections, we summarize the purpose of TWONs and the basic quality criteria.

1.1 Background

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 the emergence of 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).

However, tech companies 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,

[*] We would like to thank Achim Rettinger and Michael Heseltine for reviewing the internal draft of the paper. The report has benefited considerably from the comments.

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). In 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 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, which makes it 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.

1.2 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” of its real counterpart (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 **o**nline social networks. These models consist of two main ingredients.

First, there is a very detailed and realistic description of the network's *users*, modeling 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.

1.3 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.

1.4 Ethical Background

While we argue that TWONs have significant potential to facilitate the development of online social networks grounded in democratic and social ideals, it is also important to consider the potential negative applications of this new technology. A model designed to identify platform designs that prevent undesired outcomes can, in fact, be used to achieve the exact opposite. In other words, a functional TWON can be misused to harm the very individuals and collectives it was designed to protect.

In another deliverable, the TWON consortium has provided a detailed ethical analysis of this dilemma and outlines various governance modes for TWONs.

1.5 Quality criteria for TWONs: realism and flexibility

There is a longstanding debate about the quality criteria of theories and, in particular, formal models (Belfrage et al., 2024; Epstein, 2008). Having defined the purpose of TWONs in the previous subsections, we put forward two key quality criteria that guided the development of the TWON presented below: realism and flexibility.

First, A TWON can only be used as a convincing tool to demonstrate the effects of OSNs and to optimize and regulate OSNs if the TWON is considered *sufficiently realistic* (Belfrage et al., 2024). Otherwise, its predictions can always be disregarded as being derived from a model that is not representing the

OSN under consideration.

Nevertheless, a model cannot be a perfect replica of reality. It always needs to abstract from characteristics of the system it is supposed to represent. That is, there need to be aspects of reality that are not or only abstractly included in the TWON. The question is, thus, what aspects of reality need to be represented in a TWON and what aspects can be neglected?

In a nutshell, we argue that a TWON needs to represent all aspects of an OSN that critically affect relevant predictions of the model. All remaining aspects can be neglected. For instance, Keijzer and colleagues (2018) showed that the predictions of a classical model of social-influence dynamics depend on the model's interaction regime. Most existing models assume a so-called one-to-one interaction regime, where in a time step one agent is exerting influence on one other agent. On many OSNs like X, however, agents emit content to multiple followers at the same time. This one-to-many interaction regime, it turns out, fosters polarization, because an agent emitting content to multiple network contacts is ”pulling” these agents towards him and away from joint friends that were not affected by the content. This has a bigger potential to segregate the network that one-to-one interaction. Accordingly, if a TWON is used to study polarization and if the studied OSN has a one-to-many interaction regime, then the TWON needs to representing interaction according to this regime.

In Section 2, we reflect on a method to identify critical aspects of interaction on OSNs. In addition, we report on own analyses demonstrating that heterogeneity in user activity and its dependence on user perceived rewards needs to be included.

The second quality criterion that guided the TWON development was *flexibility*. That is, we seek to develop a model that allows us to flexibly adjust model assumptions about the design of online social networks or the behavior patterns of users. This aspect that matters for two reasons. First, we seek to be able to derive statements about various OSNs, which requires that different designs decisions need to be implemented in the TWON. Second, to demonstrate that a given characteristic of an OSN affects dynamics on the platform, it's TWON needs to be compared to a counter-factual that is identical to the TWON with one exception: the aspect under investigation needs to differ. Accordingly, we sought to develop a model that allows us to implement even unrealistic designs of OSNs.

2 How to develop a TWON?

2.1 Methodology

In Section 1.5, we defined the two central quality criteria that guided us in the development of the TWON: realism and flexibility. Developing a realistic model – a model that encompasses all aspects of the real systems affecting the system dynamics under investigation—is a significant challenge. One issue is the frequent lack of *empirical information* about whether a given aspect has meaningful effects. Indeed, one of the primary reasons for developing TWONs is the absence of this empirical information. Furthermore, it is often unclear whether an empirically observed aspect not only exists but also meaningfully impacts system dynamics. For instance, while the activity of social bots on online social networks is well documented, modeling work suggests that well-connected and highly active bots tend to influence relatively few users (Keijzer and Mäs, 2021). Therefore, the prominence of a specific characteristic in an OSN does not necessarily imply that it needs to be included in the TWON of the OSN.

Identifying relevant aspects of a system using *theoretical methods* is also problematic, a well-known issue in statistical modeling. Researchers aiming to find unbiased statistical models try to identify all variables that explain variance. When adopting a *top-down approach*, they start with a model containing all available variables and then exclude insignificant variables step-by-step. In the context of developing a TWON, this suggests first developing a model that includes all potentially meaningful aspects and then conducting analyses to test whether model predictions change when an aspect is excluded. However, for the purpose of developing a TWON, this approach is hardly feasible, as a model containing all potentially relevant aspects would be too complicated and difficult to analyze.

Alternatively, one can adopt a *bottom-up approach*. This involves starting with a very simple model and systematically adding new aspects, testing at each step whether the addition changes model predictions. If an aspect does not make a difference, it is excluded. The main disadvantage of a bottom-up approach in both statistical and theoretical modeling is that, during the process of building the final model, one studies incomplete models, which can lead to misleading conclusions. In other words, an aspect that turns out to have impact in a simple model may not have effects in the actual TWON and vice versa.

Despite these challenges, we adopted a bottom-up approach, building on the rich literature on models in the social sciences and complexity research (Mäs, 2021). We proceeded in five steps:

1. We formulate a conjecture about an aspect that may be relevant and, thus, is a candidate for inclusion in the TWON.

2. We integrate this aspect into an existing toy model from the literature. The literature on dynamics in social complex systems provides numerous well-understood toy models (Flache et al., 2017; Axelrod, 1997).

3. We test whether the model's predictions change when the aspect is included, thereby testing the conjecture within the context of the toy model.

4. If the conjecture is supported, we identify the mechanism responsible for the observed effect. This involves providing an explanation for why the aspect changes model predictions.

5. If the conjecture is supported and there is good reason to believe that the identified mechanism is also active in the TWON, we include the aspect.

This approach is not without flaws. Using toy models provides merely a good reason to include specific aspects in a TWON, as it may turn out that an aspect that changes predictions in a toy model does not matter in the final TWON or vice versa. Therefore, it is important to identify the mechanism responsible for the effect observed in the toy model and to consider whether this mechanism may also be active in a more complex model (see Step 4). To illustrate our approach, we provide an example of how we applied it in the following subsection.

2.2 Example: Does success-driven user activity matter?

Almost all models of social influence dynamics on social networks build on the assumption that individuals actors (users) are equally active (Horn et al., 2024). That is, all actors are assumed to communicate and update their opinions with the same probability. For instance, in Axelrod's seminal model of cultural dissemination, at every time point a randomly picked agent is selected for update and can be influenced by a network neighbor (Axelrod, 1997). This homogeneity assumption is highly unrealistic for the context of online social networks, where a relatively small share of users are highly active while most users contribute little (Riquelme and González-Cantergiani, 2016). However, whether this heterogeneity needs to be represented in the TWON depends on whether important model predictions change when heterogeneity in user activity is included.

Using a toy model, we tested whether heterogeneous user activity affects model predictions about the emergence of polarization. To this end, we included this heterogeneity in Axelrod's classical toy model and observed whether this extended model predicts more or less polarization compared to the original version with homogeneous activity. To be more precise, we included what we call *Success-driven user activity*, implementing that agents who have experienced successful interactions grow more active. This follows from classical learning theory, specifically social reinforcement learning (see Section 4.1). In a nutshell, it assumes that users receive social feedback from other users who consume their content by receiving likes, retweets, up-votes, and other forms of social approval. Learning theory demonstrates that behavior that receives positive feedback is reinforced and will likely be repeated (Wu et al., 2009). Accordingly, we assume that agents who experienced online activity as a success events will grow more active.

In Axelrod's model, all agents are described by a set of F features representing different cultural characteristics that are open to social influence. On each feature, agents can adopt one of Q nominal traits. At the outset of the simulated dynamics, agents are assigned a random trait on each feature. In every timestep, in Axelrod's model an agent i and one of his contacts j are randomly picked for updating. Whether there is an updating or not, depends on the cultural overlap, which includes the principle of homophily (Merton and Lazarsfeld, 1954; McPherson et al., 2001). To be more precise, it is assumed that the probability of interaction between i and j equals the share of features where the two have adopted the same trait. If they do interact, then i will adopt one of j's traits that i had not adopted before.

The process of picking two agents, determining whether they interact, and the actical influence in case of interaction is repeated until one of two possible equilibria is reached. Dynamics settle when all agents have adopted the same traits in all features, a state that can be denoted consensus or monoculture. Alternatively, a rest point is reached when the population has segregated into multiple internally homogeneous but mutually distinct subgroups. Axelrod referred to this state as ”polarization”. Our question is whether this equilibrium is reached more likely when success-driven activity is assumed.

To enable the implementation of success-driven activity, two adjustments to the original Axelrod model were necessary, as summarized in Table 1. First, we had to swap the selection of the sender and the receiver of social influence (see Step 2 in Table 1). Second, we counted the number of times an agent managed to exert influence on a network neighbor in the past (Step 5) and made the probability $P(a_{i,t} = 1)$ that an agent is selected for sending a message (Step 1) dependent on this count $s_{i,t}$.

For selecting sender and receiver and thus implementing activity heterogeneity and success-driven user activity, we used Equation 1. Due to this so-called ”roulette wheel selection”, the probability for being selected as a sender i of a certain trait Q to a network neighbor, is proportional to the relative past success $s_{i,t}$ of the agent. The degree of how strong past success $s_{i,t}$ influences the probability of being selected as a sender i can be varied and is named *success motivation* m. When $m = 0$, success does not influence activity, which implements Axelrod's model where all agents have a probability of $1/N$ to be selected for update. For $m > 0$, success-driven activity is implemented and thus influences the probability for being selected as sender i.

$$P(a_{i,t} = 1) = \frac{{s_{i,t}}^m}{\sum_{j=1}^{N} {s_{j,t}}^m} \tag{1}$$

Our findings are described in detail in a journal publication (Horn et al., 2024). In a nutshell, we observed that success-driven use activity enhances polarization. In fact, even under conditions where

<table>
<tr><th>Original Axelrod Model</th><th>Success-driven Activity</th></tr>
<tr><td colspan="2"><b>(1) Select an agent i</b></td></tr>
<tr><td>Pick an agent i with probability $1/N$. Agent i is the receiver of content</td><td>Pick an agent i with probability proportional to success $s_{i,t}$ as implemented in Equation 1. Agent i is the sender of content</td></tr>
<tr><td colspan="2"><b>(2) Select neighbor</b></td></tr>
<tr><td>Select a random neighbor j as content sender</td><td>Select a random neighbor j as content receiver</td></tr>
<tr><td colspan="2"><b>(3) Homophily</b></td></tr>
<tr><td>With probability equal to similarity between i and j execute Step (4), otherwise move to Step (1)</td><td>With probability equal to similarity between i and j execute Steps (4) and (5), otherwise move to Step (1)</td></tr>
<tr><td colspan="2"><b>(4) Social influence</b></td></tr>
<tr><td>Pick a feature on which i and j differ and let i adopt j's trait</td><td>Pick a feature on which i and j differ and let j adopt i's trait</td></tr>
<tr><td colspan="2"><b>(5) Update agent i's success count s_i</b></td></tr>
<tr><td>–</td><td>Increase i's count of successful influence events $s_{i,t}$</td></tr>
</table>

Table 1: Axelrod's original model and its version with success-driven activity.

Axelrod's model virtually always predicts that agents will eventually hold the same traits on all features, we observed that the population likely falls apart into multiple internally homogenous but mutually distinct subgroups (polarization).

The polarizing effects of success-driven user activity are particularly insightful when random perturbations are included. Following the approach of Klemm and colleagues, we introduced random perturbations by varying the parameter r. At every time step, a randomly chosen trait on a randomly picked feature is adopted by a randomly selected agent with a probability r (Klemm et al., 2003). Furthermore, we applied extra perturbations whenever the system reached equilibrium. To maintain comparability in the number of time steps, we advanced the time step count by a random value sampled from a geometric distribution with parameter r.

In Figure 1, the orange markers and corresponding line indicate that we successfully replicated the core findings of Klemm et al. (2003), who worked with Axelrod's original model. At very high perturbation rates, the dynamics result in an unordered state with trait distributions resembling randomness. However, at intermediate levels of perturbation, the system exhibits ongoing fusion and fission of cultural clusters, leading to polarization. When success-driven activity is introduced, we observe a similar pattern, but with much stronger polarization. As shown by the blue markers of Figure 1, the runs are characterized by prolonged phases of more intense polarization than under homogeneous activity.

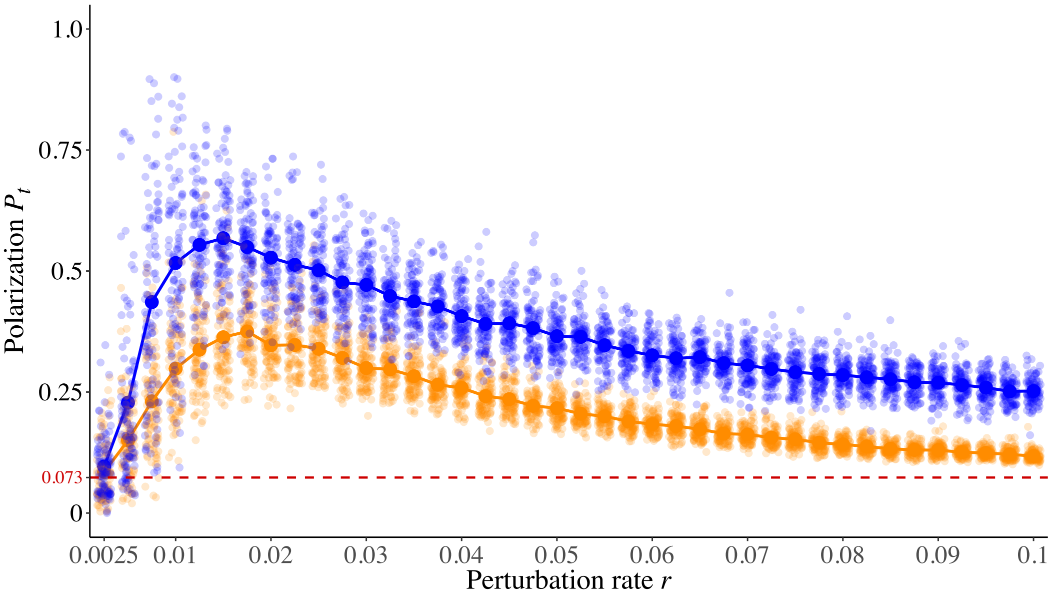

Figure 1: Effect of different perturbation rates on polarization P_t in equilibrium. P_t is the variance of the pairwise cultural dissimilarities between all $N \times N$ pairs of agents (Horn et al., 2024). We conducted 100 independent simulation runs per experimental treatment, always assuming a population size of $N = 100$, number F of features = 5, number Q of traits = 20; and a degree of 8 for all agents. The original Axelrod model $(m = 0)$ is shown in orange. The success-driven activity $(m = 1)$ is shown in blue. Dots linked with lines show treatment averages.

Thus, Figure 1 shows that according to Axelrod's toy model, polarization is stronger when activity heterogeneity in terms of success-driven activity is included. This is an important reason to include this aspect also in the TWON. However, we argue that a *second reason* is required: there must be a theoretical argument suggesting that the same effect should emerge in the TWON. To this end, the mechanism generating the effect in the toy model needs to be identified and it must be plausible that this mechanism is also active in the TWON.

The *mechanism* contributing to polarization in the toy model has two ingredients. First, success-driven activity implies a rich-get-richer dynamic: individuals who achieve success are more frequently activated, which enhances their probability of influencing others and further increasing their success.

The result is a highly skewed distribution of success, where a few agents are highly successful and many agents have experienced few successful interactions. Second, the highly successful and, as a consequence, highly active agents tear apart the population into distinct subgroups, since they exert very frequent influence on their neighborhoods. This local convergence of traits contributes to global divergence and polarization. We failed to come up with an argument suggesting that one of the two ingredients is an artifact of the toy model and should be inactive in the TWON. Accordingly, we deem it important to include success-driven activity in the TWON.

3 Overview over the TWON

3.1 Conceptual model

Before, we detail the TWON's theoretical assumptions and their formal implementation, we describe here the general structure of the formal model. Next (Section 3.2), we describe the nuts and bolts of the formal implementation.

In a nutshell, the TWON consists of five key building blocks, as visualized in Figure 3.1. First, there is what we denote the **technical infrastructure** and detail in Section 7. The technical infrastructure includes both hardware and software. The software encompasses every task of coding, that is including all aspects of the other building blocks: all details of platform model, user model, their interaction, and evaluating outcome measures. The hardware consists of essential computing resources such as maintaining servers, computing power, but also simply the data storage.

The second building block is the **platform model**, which represents the core functionality and rules governing the online social network represented by the TWON (see Section 5). This includes platform affordances such as the presence of like and dislike buttons, the ability to comment or share, or character limits. Furthermore there are algorithms for content ranking, influencing the content each user is exposed to. Additionally, hate speech detection can inform content moderation tools like banning, shadowbanning, or human content moderation can be simulated by limiting the fraction of reviewed content.

The third building block is the **user model**, which contains all assumptions about human behavior on the platform and which we describe in Section 4. It includes all assumptions determining when and how long users interact with the platform, which content they choose to consume and their decision making process regarding engagement with the platform's content. Furthermore, it accounts for user-generated content, modeling both when and what users decide to write.

Fourth, the **user-platform interaction** examines the dynamic relationship between users and the

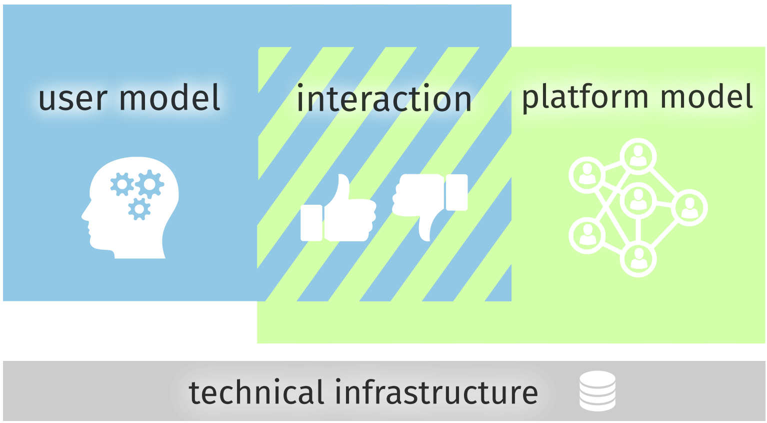

Figure 2: Conceptual model of the TWON. Built on top of the technical infrastructure, there are three elements: the user model encompassing psychological assumptions about human behavior, the platform model capturing all platform affordances and third, the emerging interaction between the two.

online social network. This mainly involves how design choices influence user behavior, especially the macro consequences of interaction between users. To be clear, this building block does not contain any additional assumptions, rather it results from the interaction between users and the interaction between users and the online social network.

Finally, the framework includes **outcome measures**, which evaluate the overall quality of discourse within the simulated environment. These metrics include indicators of debate quality, levels of polarization, toxicity, and the formation of filter bubbles. Outcome measures are described in detail in Deliverables D5.1 and D5.2.

In a nutshell, the TWON outcome measures are based on two aspects of the model. First, we analyze the structure of the emerging communication network. While, we can also manipulate the structure of the network of who can in principle communicate with whom (see Section 5), agent decisions and algorithms implemented on the platform model determine who is actually communicating with whom. The structure of this network may be very cohesive or it may be characterized by clustering and segregation, which is a key outcome variable (Moody and White, 2003).

Second, we analyze the content generated by the virtual users of the TWON using established methods (Mohammad et al., 2018; Van Hee et al., 2018; Basile et al., 2019; Zampieri et al., 2019; Rosenthal et al., 2017). The TWON is a network of large language models (LLMs) generating, sharing, and responding to content. Accordingly, the outcome variables of our experiment will be based on the characteristics of the content generated by the LLMs. We choose outcome measures that allow us to gauge the debate quality arising in the TWON (see Deliverable D5.1). At the moment, we are measuring for every piece of content the following characteristics

- Hatefulness,
- Offensiveness,
- irony
- Sentiment (negative, neutral, positive),
- Emotions (anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise, trust),
- Topics (arts and culture, business and entrepreneurs, celebrity and pop culture, diaries and daily life, family, fashion and style, film tv and video, fitness and health, food and dining, gaming, learning and educational, music, news and social concern, other hobbies, relationships, science and technology, sports, travel and adventure, youth and student life)

Summary 2

The TWON consist of five building blocks:

1. **The user model** contains all model assumptions about the behavior of users, including their decision to interact with the OSN, and to consume, evaluate, share, or create content
2. The **platform model** is the set of assumptions describing the design of the OSN, including assumptions about platform affordances and algorithms governing what content users are exposed to.
3. **The users-platform interaction** arises from the behavior of users and platform design aspects. For instance, when
4. **Outcome variables** capturing the dynamics generated by the TWON.
5. **Technical infrastructure**

3.2 Computational model

Figure 3 provides an overview over how the dynamics emerging from user decisions and platform responses.

We model time as a sequence of discrete time steps t, called *rounds*. The exact duration a round represents is not clearly defined, but we interpret it as a period of no more than a few minutes. The relationship between a round and actual time depends on parameters such as the size of agents' resource budgets or the resource reduction resulting from specific behaviors.

Before a round starts (see the violet boxes in Figure 3), the TWON decides which agents will be activated to engage with the online social network. Agents that were inactive in the previous round $t-1$ may decide to log on, while agents who were active may choose to stop interacting with the platform. In Figure 3, for instance, Agent 4 turns inactive at Round 2, while Agent 3 is activated before Round 2.

Next, all agents active in a round interact with the platform synchronously. While calculations for different agents can be performed sequentially, the synchronous schedule ensures that agents active in a given round base their behavior on the system's state at the very beginning of that round. Only when the round has ended are agents exposed to changes in the system resulting from other agents' actions in the same round. This implies that agents cannot exchange messages during a round. For instance, while agent i might send a message to agent j in one round, it will only appear in j's feed in the subsequent round. This is an approximation intended to save on computation of feeds. With this approximation, feeds only need to be calculated once per active agent per round instead of continuously. The latter can be achieved though by setting the duration of a round to a short time.

The activities of an agent are calculated as follows: First, TWON updates central variables for the active agent. For instance, it compiles a list of all messages that other agents have sent to the focal agent in the past and sorts these messages according to a specified ranking algorithm. Next, the agent makes a series of decisions. The agent may decide to create and share content, consume a share of the messages received from others, and decide whether and how to react to each message. Each of these actions is costly in terms of time and energy, so the agent's remaining budget is updated after every decision. In case an action consumes more time than is left in the current round, the agent remains committed to this action as many rounds as required.

When all agents active during a given round have finished engaging with the platform, the round ends, and the next round starts. Unlike many abstract models (see e.g. Section 2.2), the TWON does not have a clearly defined rest point, a state of equilibrium where further rounds do not change the state of the system. Accordingly, the described process is iterated for an initially defined number of rounds.

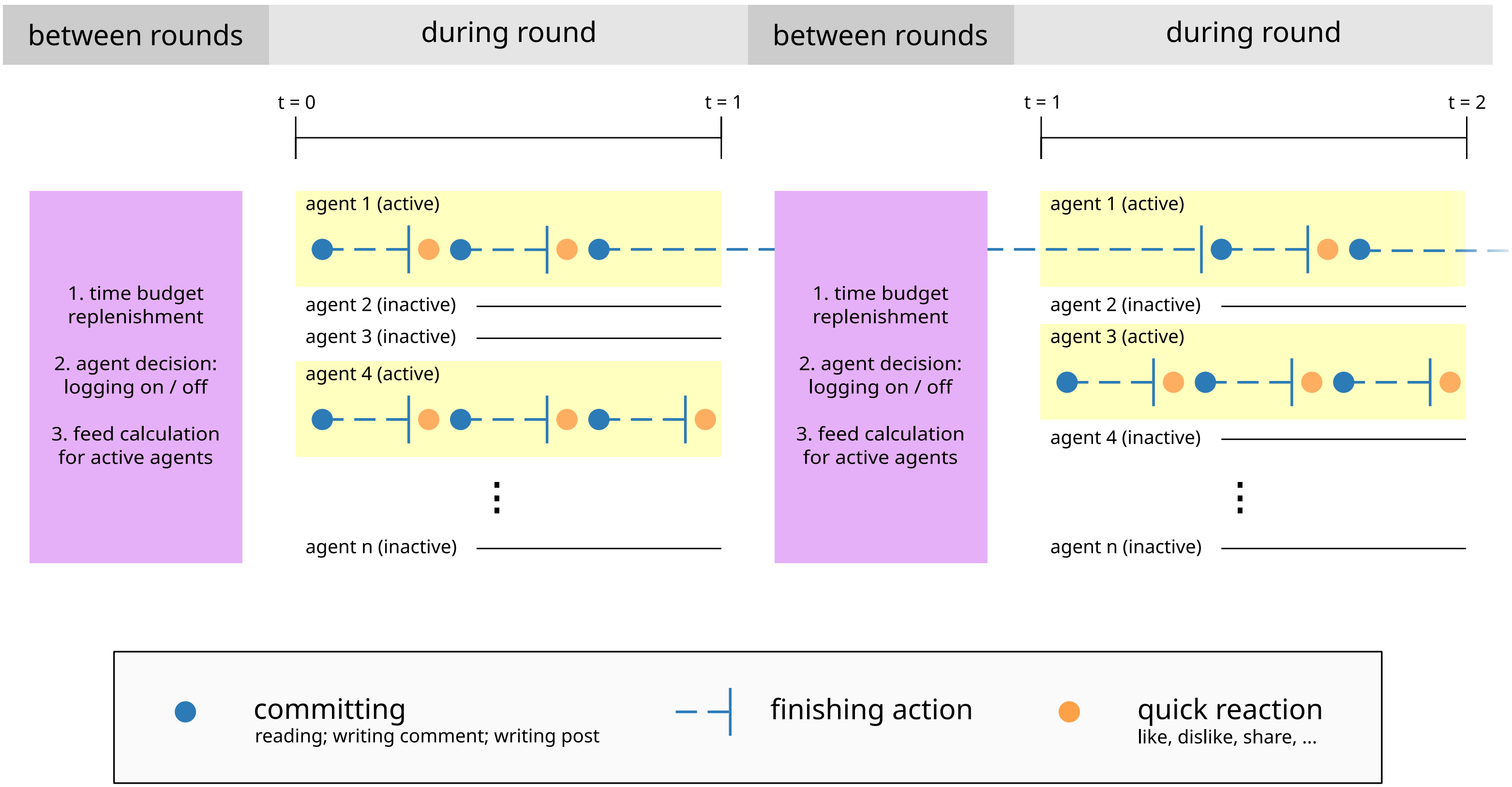

Figure 3: Sketch of the computational model. Time is a sequence of discrete time steps, called rounds. During rounds, agents who are active on the platform can decide to engage with content or produce their own. Once they decide to engage in such a time costly action (blue circle), they are committed to that action and unable to do anything else until they finish (blue vertical line). They also have the option to perform quick reactions (orange circles) which require far fewer resources. In between rounds (violet boxes), each agent's budget is replenished, it is determined which agent will be active and inactive during the next round and all active agent's feeds are calculated.

4 The user model

4.1 Theoretical foundation

Much of the behavior exhibited by TWON's virtual users is not predefined by explicit assumptions; instead, it is derived from empirical data. Specifically, the content communicated by these users is generated by a Large Language Model (LLM). An LLM is a sophisticated AI system trained on vast amounts of human text. It learns to understand and produce language that closely mimics human communication. By analyzing patterns, structures, and contexts within the training data, the LLM can generate coherent and contextually appropriate content, making the interactions with virtual users appear more natural and human-like. The great advantage of this approach is that the communication generated by the TWON is highly realistic - much more realistic than alternative approaches.

However, in the TWON there are also various forms of user behavior that have not been derived from empirical data. One reason is that for many decisions, data is not available. For instance, it is possible to observe that a user has responded to the content shared by another user. We usually have no data about users who have consumed the same piece of content but decided to not react. Accordingly, it is hardly possible to infer behavior patterns from data. Another reason why using purely empirical approaches can be problematic is that often the resulting model mimicking the behavior of humans is a black box, which makes it very hard to manipulate user behavior for experiments.

In the remainder of this subsection, we describe the theoretical assumptions underlying all user model assumptions that have not been adopted from empirical data.

In the social sciences, there is a longstanding debate about theories that describe, explain, and predict human behavior. While this scholarly discussion is marked by disagreement, most contributors concur on two fundamental components of behavioral theories: preferences and constraints.

First, it is generally assumed that human decision-makers have *preferences*—that is, their behavior is shaped by their motives and desires. Humans can rank the outcomes of their actions based on how desirable they find them. While theories differ significantly regarding the stability of preferences and the extent to which behavior aligns with them, preferences always play a role in shaping behavior according to the literature.

There is also debate about which preferences are relevant. The narrow version of the rational-choice approach, for instance, assumes that decision-makers are egoistic and consider only economic outcomes (Opp, 1999). Broader versions, however, adopt a less restrictive view (Fehr and Schmidt, 1999). In the context of online social networks, Lindenberg's Social Production Function Theory provides useful insights (Ormel et al., 1999; Lindenberg, 1986). This theory posits that human well-being arises from achieving two instrumental goals: physical well-being and social well-being—both of which are relevant in the context of TWONs.

Physical well-being is affected by online behavior when users can obtain financial rewards. For example, YouTube allows content creators to monetize their accounts once they generate sufficient engagement. Their income depends on how frequently users interact with their content—specifically, the more views a YouTube video receives alongside an advertisement, the higher the creator's earnings. Social well-being, on the other hand, is enhanced when individuals receive validation from others. As social beings, humans derive satisfaction from knowing that their actions are appreciated. Positive feedback, such as likes, comments, or other forms of engagement, contributes to this sense of social well-being.

Accordingly, we assume that users seek to engage in behavior that elicits feedback—whether positive or negative—from others. The more feedback an individual receives, the more rewarding the behavior is perceived to be.

The second key component of behavioral theories is constraints. Behavior is always associated with costs, as engaging in online activities—such as logging in, creating content, reading posts, and responding—requires time and effort. We therefore assume that users have limited resources in terms of time and energy, which constrain their online activity.

While there is little debate about the importance of preferences and constraints, theories differ significantly in their assumptions about *decision rules*—the mechanisms by which individuals translate their preferences and constraints into actual behavior.

We implemented a decision rule that implements what is typically referred to as Bounded Rationality (Simon, 1990). While classical rational-choice approaches assume that decisions are perfectly consistent with preferences and restrictions, a bounded rationality includes that there may be deviations from rationality. We included four aspects of bounded rationality in our model:

1. We assume that agents are *backwards looking* rather than forward looking (Heckathorn, 1996). While perfectly rational agents anticipate the future consequences of their actions and determine the utility they generate, we assume that agents consider how rewarding different behavioral options were in the past and tend to choose options that were experienced as most rewarding.

2. Our agents are *myopic*, in that they do not predict future consequences. A rational actor, for instance, would form predictions about how other users will respond to one's own behavior and how these responses affect his own outcomes. Rather, our agents tend to choose a behavior that promises a high reward given their past experiences.

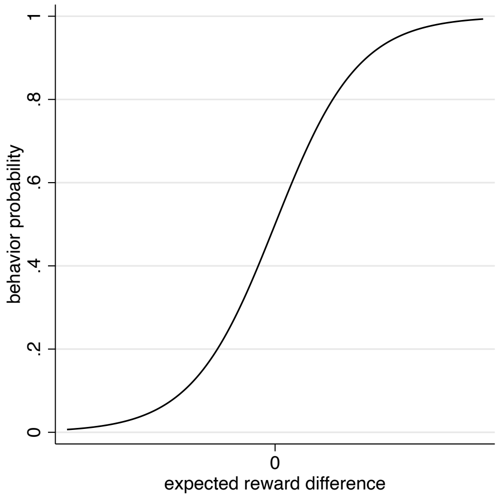

Figure 4: The logistic function as an implementation of the myopic best-response model

3. We include random deviations from otherwise preferred options (Mäs and Helbing, 2020; Mäs and Nax, 2016).

4. We allow agents to switch between different modes of behavior. When login in, for instance, we include the assumption of decreasing marginal reward. However, once they have spend considerable time on t platform, we include that agents may experience it hard to stop consuming content, a form of addictive behavior.

The *decision rule* used by the TWON's agents is the so-called ”myopic best response rule”. In the most common case where an agent decides between two options, this decision rule translates the difference between the expected rewards from the two options into a probability that one of the options is selected. Figure 4 illustrates how this is implemented with the logistic function. When the expected reward of a behavior such as sharing content exceeds the expected reward of not sharing the content, the user most likely will choose to share the content.

Summary 3

Central assumptions of user behavior

1. **Preferences:** We assume that user behavior is reward driven and that user experience feedback from other users as rewarding. For instance, users who have received many likes after posting content, experience this as rewarding and tend to repeat the behavior.
2. **constraints**: Users have a limited cbudget of time and energy. Whenever they are active, this budget decreases.
3. **Decision rule**: User apply the myopic best-response rule.

4.2 Resource budget

Implementing constraints, we assume that every agent is described by two personal budgets. First, an agent's general resource budget $b^g_{i,t}$ $(0 \le b^g_{i,t} \le 1)$ captures the energy and time available to agent i at a given point in time t. This resource budget decreases whenever an agent engages in an activity by an amount dependent on the given activity. However, this budget also regrows at every round by an amount s $(s < 1)$. It is also possible to define for every agent i a personal maximal value of $b_i^{g,max}$ ($(b_i^{g,max} \ge b^g_{i,t}$), to implement that agents differ in the resources available for online activity. Second, there is also a maximal round-resource budget determining how much energy and time an agent can spend during a single round $b^r_{i,t}$ ($b^r_{i,t} \le b^g_{i,t}$). When this budget is used up, an agent cannot engage in further activities in a given round independent of how much general budget is still available.

4.3 User activation

Each user can be in one of two states, online or offline. At each time step, an offline agent can login into the platform, and an online agent can logoff.

In our TWON, we assume that each binary decision follows the Myopic Best Response Model (see Section 4.1), where users compare the rewards of their two possible choices based on past experiences.

4.3.1 The decision to log on

In the following we will use the Myopic Best Response Model to find the probability of a user to log in during a certain period of time, for example $\Delta t = 1$ h. However, in the TWON we use a more granular

Symbol	Interpretation	Default Value	Equation
$b^{g}_{i,t}$	General resource budget available to agent i at time point t	-	10
$b^{g,max}_{i}$	Maximal amount of resources agent i can possess	1	10
s	Strength of resource regrow per time point	0.01	-
i	index of the user sending a message	-	
j	index of the user receiving a message	-	
m_i	motivation of agent i to be active	-	
e_i	Engagment of user i	-	
s_i	Success of user i	-	
s^k	Success of post k	-	

Table 2: List of symbols concerning agents' resource budgets

time discretization, namely the length of a round $\delta t < \Delta t$, and therefore we need to rescale the logon probability accordingly. For doing so we will assume that logging on is a memoryless Poissonian process:

$$P_{\delta t} = 1 - e^{-P_{\Delta t}\frac{\delta t}{\Delta t}} \simeq P_{\Delta t}\frac{\delta t}{\Delta t}. \tag{2}$$

Where we assume that the probability of going online in a fixed interval of time Δt follows a logistic function:

$$P\left(\text{ON}\right) = \mathcal{L}\left(\Delta\text{EU}_{\text{ON}}, \beta_{\text{ON}}, 0\right), \tag{3}$$

where $\Delta\text{EU}_{\text{ON}} = \text{EU}\left(\text{ON}\right) - \text{EU}\left(\text{not ON}\right)$ is the net reward, or expected utility, of logging on, and β controls the response sensitivity to the net reward. In the first iteration of the TWON we assume $\beta_{\text{ON}} = 1.5$, as estimated in Mäs and Nax (2016), furthermore, we assume that there is no offset $X_{\text{ON}} = 0$, meaning that if logging on has the same expected reward of not logging on, the probability of logging on in the interval of time Δt of reference is $\frac{1}{2}$.

The reward function is a linear combination of four rewards/cost functions:

- Cost for spending time online, scaling with the expected amount of time a user will spend online $\mathbb{E}[T]$, $u_t(\mathbb{E}\left(T\right))$
- Reward from personal value, usually entertainment, during $\mathbb{E}[T]$ time online, $u_v(\mathbb{E}\left(V\right)|\mathbb{E}\left(T\right))$
- Reward from social feedback during $\mathbb{E}[T]$ time online, $u_f(\mathbb{E}\left(F\right)|\mathbb{E}\left(T\right))$;

- Reward from checking the social media after a notification (FOMO), $u_{\text{FOMO}}\left(t,\bar{t}\right)$:

$$\text{EU}_{\text{ON}}\left(\mathbb{E}\left(T\right)\right)=w_{t,\text{ON}}u_t(\mathbb{E}\left(T\right))+w_{e,\text{ON}}u_v(\mathbb{E}\left(V\right)|\mathbb{E}\left(T\right))+w_{f,\text{ON}}u_f(\mathbb{E}\left(F\right)|\mathbb{E}\left(T\right))+u_{\text{FOMO}}\left(t,\bar{t}\right), \quad (4)$$

where $\mathbb{E}\left(t\right)$ is the amount of time a user is expecting to spend online, $\mathbb{E}\left(V\right)|\mathbb{E}\left(T\right)$ the amount of personal value the user is expecting to gain given the amount of time it's expecting to spend online, and $\mathbb{E}\left(F\right)|\mathbb{E}\left(T\right)$ the amount of social feedback a user is expecting to gain given the amount of time it's expecting to spend online. For a detailed description, and explanation of these three quantities, see 4.3.2.

4.3.2 Managing expectations

To model these expectations realistically, while maintaining computational tractability, we imposed several key assumptions:

1. **Weighted average of past experiences**: We assume a weighted average with an exponential discount to evaluate the amount of time spent online, the past personal entertainment, and the past social feedback, this effect is sometimes referred to as ”shadow of the past” **?**
2. **Time Expectation Bias**: For users, their expected online time will be affected by an optimism bias (Weinstein, 1980), where they will tend to disregard past experiences and they will decide wether to log in according to a smaller expected amount of time they will spend online. For tractability, in this first iteration of TWON, we will assume that the expected amount of time under optimism bias is the amount of time needed to consume one piece of content, τ_C, as it will be described below, sec. **??**

Let us start from the expected amount of time a user will spend online: $\mathbb{E}\left(t\right)$. Here there are two extreme scenarios under which a user could behave. On the one end it could behave rationally, and in its expectation of how much time it will spend online it will simply average across past experiences weighing them according to the shadow of the past principle, or it could behave in a maximally optimistic fashion, by assuming it will go online to take a short break, and consume a single piece of content, so for a time τ_C. We capture this behavior with a parameter α. When $\alpha=0$ the user behaves in a rational fashion, and when $\alpha=1$ in an optimistic one:

$$\mathbb{E}\left(t\right)=\alpha\tau_C+(1-\alpha)\frac{\sum_j^K e^{-\gamma\left(t-t^{(j)}\right)}T_j}{\sum_j e^{-\gamma\left(t-t^{(j)}\right)}}. \quad (5)$$

Once the user has an estimation of how much time it will spend online, it can use it to evaluate how much reward it will get from personal value, and from social feedback, for an expected amount of time $\mathbb{E}\left(T\right)$ spent online:

$$\mathbb{E}\left(X\right)|\mathbb{E}\left(T\right) = \mathbb{E}\left(T\right)\frac{\sum_k W\left(t, t^{(k)}\right)\frac{X^{(k)}}{\Delta T^{(k)}}}{\sum_k W\left(t, t^{(k)}\right)\frac{1}{\Delta T^{(k)}}}. \tag{6}$$

The sum over k represents the sum over all of the user previous experiences online; $X^{(k)}$ is either the amount of personal value or of feedback received during session k; $T^{(k)}$ is the amount of time the user spent online during the session k; $t^{(k)}$ is the time when session k began; and $W\left(t, t^{(k)}\right)$ is the weight of that experience, exponentially decreasing for past experiences, here we assume: $W\left(t, t^{(k)}\right) = e^{-\gamma\left(t-t^{(k)}\right)}$, with $\gamma = 1$; the fraction $\frac{X^{(k)}}{T^{(k)}}$ represent the personal value/feedback density: the same amount of X experienced in session k will have a higher weight if session k was short. The denominator is simply a normalization factor. The expected personal value, or feedback, is therefore interpretable as the expected amount of time spent online multiplied by the expected intensity of reward.

For future readability we define:

$$\widetilde{W}^{(k)} = \frac{W\left(t, t^{(k)}\right)}{\Delta T^{(k)}}; \tag{7}$$

this allows us to express Eq. 8 as the more readable:

$$\mathbb{E}\left(X\right)|\mathbb{E}\left(T\right) = \mathbb{E}\left(t\right)\frac{\sum_k X^{(k)}\widetilde{W}^{(k)}}{\sum_k \widetilde{W}^{(k)}}. \tag{8}$$

[??? another possible mechanism to predict success/entertaiment is linear extrapolation. If yesterday I had twice as much fun as the day before, today I will have even more fun. This consideration might be relevant for monetization in particular, since it's a much more ”rational” source of reward.]

4.3.3 Time cost function

As described in 4.1 each user is characterized by a time-dependent budget $b_{i,t}$ which gets updated according to the user's activity. The time utility function is a function of these two parameters, a function such that spending a small amount of time online leads to an arbitrary small cost ($u_t(0) = 0$), and that it goes to infinity when the expected time spent online approaches the current time budget of an user:

$$u_t\left(\mathbb{E}\left(t\right), b_{i,t}\right) = \begin{cases} \frac{1}{\left(\mathbb{E}(t) - b_{i,t}\right)} + \frac{1}{b_{i,t}} & \text{if } \mathbb{E}\left(t\right) < b_{i,t} \\ -\infty & \text{otherwise} \end{cases}. \tag{9}$$

4.3.4 Social feedback reward function

To characterize the social feedback reward function we assume a diminished return on the support, or backlash that the user is expecting to receive from other users. The social feedback is denoted by the letter F, and is composed of several contributions F_j. In a multi-topic discussion, for example, a user can belong to a filter bubble where topic A is often discussed and positively evaluated, and reshared, while topic B is frowned upon, disliked or hated. The user will receive positive feedback about its activity involving topic A leading to a positive F_A, and negative feedback about its activity involving topic B, with negative value of F_B.

One could separate the feedback from other users into two categories *immediate* and *delayed* feedback. The former is feedback about content generated, or content shared that has been received in the same session where such a piece of content was generated, and the latter is feedback received during future sessions, or between sessions. While it's trivial to attribute the immediate feedback to the session when both the content and the feedback was generated, the decision of when to attribute the *delayed* feedback is more arbitrary, and more dependent on the psychology of the user. For tractability in this TWON prototype we will assume that each feedback is altering the amount of success the user experienced in the session where that piece of content was generated.

Given a certain expected social feedback $\mathbb{E}(F)$, we assume a diminished return functional form for its expected utility, with the possibility of a different scaling for negative feedback if people are more loss averse:

$$u_f(\mathbb{E}(F)) = \begin{cases} \log(1+\mathbb{E}(F)) & \text{if } \mathbb{E}(F) \geq 0 \\ -\lambda_F \log(1-\mathbb{E}(F)) & \text{if } \mathbb{E}(F) < 0 \end{cases} \tag{10}$$

In the first TWON prototype we will assume, for tractability, $\lambda_F = 1$. A value $\lambda_F > 1$ represents a heightened sensitivity to negative social feedback and potential reputational costs.

This approach comes with a main simplification: there is a single topic of discussion, or equivalently that the positive feedback resulting from supporting topic A and the negative feedback about supporting topic B are averaged together. Another possible approach is to classify each post in topic, and evaluate the success of each topic independently. In this extended framework instead of $u_f(\mathbb{E}(F))$ one should consider $u_f(\overrightarrow{\mathbb{E}(F)})$:

$$u_f(\overrightarrow{\mathbb{E}(F_k)}) = \sum_l w_{f,l} u_f(\mathbb{E}(F_l)), \tag{11}$$

where $w_{f,l}$ represents the importance of topic l for user i, and $\mathbb{E}(F_l)$ the feedback that user i expecting to receive about topic l.

For tractability, for this TWON prototype, we will merge all the social feedback into a single contribution F.

Personal personal value reward function

To characterize the personal value, which will mainly driven by the entertainment of a user, we need to approximate the dopaminergic response of the users. Since a precise description of the dopaminergic system of each agent would go beyond the scope of the TWON prototype, we will make a series of assumption, trying to capture some of the main aspects of the dynamic. Following the steps of the social feedback we will assume that the expected engagement of a user will follow a diminished return:

$$u_v(\mathbb{E}(V)) = \begin{cases} \log(1+\mathbb{E}(V)) & \text{if } \mathbb{E}(V) \geq 0 \\ -\lambda_V \log(1-\mathbb{E}(V)) & \text{if } \mathbb{E}(V) < 0, \end{cases} \tag{12}$$

where, now a negative $\mathbb{E}(V)$ represents user expecting to have a net negative amount of personal value based on past experiences. A negative $\mathbb{E}(V)$ is a consequence of being exposed to content the user dislikes, or that it is disturbed by.

The important distinction about personal value and social feedback is the fact that the former is a direct consequence of the timing and pattern of the content consumption, while the second is a direct consequence of how others reacted to our content.

4.3.5 Other factors

In this first TWON prototype we included what we believe are the more relevant contributors to personal motivation. We are well aware that user's are driven by several other factors, among which:

1. Opportunity cost of other activities. Partially covered by the time cost, but different users can have different opportunity costs, it the opportunity cost could be dependent on the time of the day/week, and on the available offer of offline activities.

2. Monetization. Some platforms offer monetary incentive for users who generate popular content, this increases the motivation to go online to generate a certain type of ”palatable” content.

Other mechanisms, such as the lack of a critical mass of active users, could be incorporated in a more nuanced personal value function. Cognitive load, attention cost, current mental state, could shape both the probability of going online, the amount of time spent online, and the content diet consumed while online. But carefully quantifying, and validating them, goes beyond the scope of this TWON prototype.

4.3.6 pop-up notifications

Online social networks have an opt-out setting that informs you when one of your posts has been liked, or shared, or commented, or when you received a personal message (a mechanism we decided not to include in this TWON prototype). The main effect of these communications is to stimulate the dopaminergic system of users increasing the expected utility from an immediate login. This mechanism can be included in several ways, we chose to implement it in terms of FOMO, (Fear Of Missing Out), creating a negative utility associated to the decision of not logging in after receiving a notification. For tractability purposes, we will assume that this negative utility will decrease exponentially with time, and that a second pop-up notifications while online only resets it at its maximum value:

$$u_{\mathsf{FOMO}}\left(t,\bar{t}\right) = w_{\mathsf{FOMO}} e^{-\gamma_{\mathsf{FOMO}}\left(t-\bar{t}\right)}. \tag{13}$$

where $\bar{t}$ is the time of the last notification. In the following, for tractability we will assume that each user have the same value of γ_{FOMO}, and that the FOMO-utility has the same weight for each user, while in reality there is a high variability on how strong the response of pop-up notification is across users. Furthermore, its strength is prone to the saturation effects, where the value of w_{FOMO} decreases in case of excessive notifications.

[??? possible intervention: different weight of notification accordingly to the difference between users. if L likes a R comment R gets/not get a notification]

Including this last contribution into Eq. **??** we obtain:

$$\begin{aligned}\Delta \mathsf{EU}_{\mathsf{on}} =& w_t \left[u_{\mathsf{t}}(\mathbb{E}\left(T\right)) - u_{\mathsf{t}}(0)\right] \\ &+ w_e \left[u_E(\mathbb{E}\left(E\right), \mathbb{E}\left(T\right)) - u_E(\mathbb{E}\left(E\right), 0)\right] \\ &+ w_f \left[u_F(\mathbb{E}\left(F\right), \mathbb{E}\left(T\right)) - u_F(\mathbb{E}\left(F\right), 0)\right] \\ &+ u_{\mathsf{FOMO}}\left(t,\bar{t}\right).\end{aligned} \tag{14}$$

4.4 User platform actions, and deactivation (owner: Fabio)

Once a user logs into the platform, the user is considered *Active*. At this point each user might engage in a sequence of actions that can be broadly categorized into two categories, according on their visibility:

private and *public* actions.

Private actions are actions where the user can hide behind a certain degree of anonymity. A classical example is reporting, completely anonymous, blocking, visible only by the user who has been blocked, and, most relevant for our TWON liking and disliking. While liking and disliking are technically a public action in the majority of the online social network, their visibility is much lower than other possible reactions such as forwarding or commenting. Therefore, users are generally not held accountable, nor rewarded for which pieces of content they like or dislike. The decision of taking a *private* action solely depends on the user mind state, its time availability, the interest in a certain topic, and on its opinion. While, the decision of taking a *public* action also depends on the social feedback the user is expecting to receive from such action.

While the user is active, it will loop between a list of possible actions available in the platform. If it decides not to engage in any of the available actions it will log off. The actions it can decide to do can be classified in three groups: **Content Creation** (sec. **??**) and **Content Reaction** (sec. **??**), In principle, a third category is also possible, but it's not included in the TWON prototype: **being idle**.

In the following we will include an exhaustive list of mechanisms, while all of them might play a role for the metrics discussed in sec. 6, not all of them are implemented in our prototype.

Once a user will log on, it will repeat the following until it logs off.

1. **Proactive Content Creation** decide whether to generate a novel piece of content (*public*)

2. **Content Reaction** decide wether to read the next post suggested by the recommending algorithm. Conditional to the user reading the piece of content, it will:

    - decide if liking/disliking it (*private*)
    - decide if commenting it (*public*)
    - decide if forwarding it (*public*)
    - decide if reporting the user (*private*)
    - decide if blocking the user (*private*)

3. **Staying idle** the user will decide wether to wait online without doing anything. Relevant for future developements where the psychological mecanism of ”waiting to see what they will answer” will be implemented, currently, without the perspect of the future rewards implemented, the decision of a user will always be not to stay idle.

Each of these decision will be take according to the same logic of the myopic best response model used for the log-in 4.3.1, where the reward of doing an action is compared against the reward of not doing an action.

4.4.1 New content generation (DONE)

The decision of generating a piece of content proactively, namely not as a response to another piece of content is generally influenced by several factors, the feeling of urgency of communicating a concept, the time available, the interest in a topic, the expected success the post will have, and the expected monetary compensation for generating such piece of content. In this first TWON prototype we will focus on two of the main mechanisms: expected feedback/validation and time constraints.

The expected reward generated by the social feedback on a single piece of content, is a similar logarithmic diminished return as in Eq. 10 of the weighted average of the past feedback received on posts, comments, and forwarding of content on a similar topics, or in case of the single topic approximation, posts, comments and forwardings in general:

$$\mathbb{E}\left(F\right)=\frac{\sum_{l}\widetilde{W}^{(l)}F_{l}}{\sum_{l}\widetilde{W}^{(l)}}, \tag{15}$$

where $\sum_l$ sums over all the previous relevant posts, comments, and forwards, and $\widetilde{W}^{(l)}$ is the weight described in Eq. 7.

The net feedback reward for writing a piece of content k can be written as:

$$\Delta u_{G,f}\left(k\right)=u_{f}\left[F+\mathbb{E}\left(F\left(k\right)\right)\right]-u_{f}\left[F\right] \tag{16}$$

, where u_f is defined in Eq. **??**

The positive expected reward from feedback is balanced by the negative expected cost of spending time to generate a piece of content. For tractability we will assume that each piece of content was generated in the same amount of time τ_G, the cost for generating a piece of content, therefore can be calculated according to Eq. 9:

$$\begin{aligned}\Delta u_{t}\left(\tau_{G}\right)&=u_{t}\left(\tau_{G},b_{i,t}\right)-u_{t}\left(0,b_{i,t}\right)\\&=\frac{-\tau_{G}}{b_{i,t}\left(b_{i,t}-\tau_{G}\right)}\end{aligned} \tag{17}$$

Another contribution to the negative expected cost is the opportunity cost, the same time the user will spend writing a post, or creating a picture could have been spent consuming content experiencing personal value. For this TWON prototype we will assume that such costs are already included in Eq. 17.

The total reward from writing a post can be written as:

$$\mathsf{EU}_G = w_{G,f} \Delta u_f \left(\mathbb{E} \left(F \right) \right) + w_{G,t} \Delta u_t \left(\tau_G \right). \tag{18}$$

The probability of writing a post is therefore given by:

$$P_{\mathsf{G}} = \mathcal{L} \left(\mathsf{EU}_G, \beta_G, 0 \right). \tag{19}$$

4.4.2 Content consumption (OWNER: FABIO)

When a human encounters a new piece of content, either a post, or a comment on a post, their choice to engage with it is heavily influenced by their cognitive state and accumulated fatigue. The brain experiences mental fatigue from processing information, making users progressively less likely to engage with complex or lengthy content as their session continues. This fatigue effect interacts closely with time constraints – users constantly, evaluate how much free time they have available and weights it against their accumulated time already spent consuming content, similar to a mental budget that depletes throughout their session. The initial decision to engage often happens within seconds, based on rapid assessment of content characteristics. Users quickly scan titles, thumbnails, or opening lines to gauge relevance to their interests, while simultaneously making snap judgments about the required time commitment based on content length indicators (video duration, article length, etc.) and format (text requires more active engagement than video, for instance). The content source plays a crucial role – users often have ”automatic yes” sources, like favorite creators, where they'll immediately engage regardless of other factors. Two powerful psychological forces often override these rational assessments: FOMO (Fear of Missing Out), which creates anxiety about potentially missing important or trending content, and ”scroll momentum,” where users fall into a pattern of continuous scrolling that makes them less likely to break the rhythm to engage deeply with any particular piece of content. This scroll momentum acts like a behavioral inertia that must be overcome for interrupting the unquestioned stream of content consumption.

We model the decision of consuming another piece of content using the same myopic best response framework. Where the main contribution to the reward are the time needed to consume a piece of content and the expected personal engagement resulting from consuming it. Further psychological mechanisms that can be included as mentioned before are FOMO, as additional negative reward for not consuming that content, scroll momentum as a increased probability of keep consuming a piece of content dependent on how much content has been consumed consecutively, mental load can be modeled in a similar way to time budget.

The reward from personal value stemming from a user consuming a piece of content k, without considering the abovementioned FOMO mechanisms can be modeled using the same equation used for the login decision, Eq. **??**. Therefore the net reward can be written as:

$$\Delta u_{C,v}\left(k\right) = u_v\left[V + \mathbb{E}\left(V\left(k\right)\right)\right] - u_v\left[V\right] \tag{20}$$

If a user is fully driven by FOMO that can be captured by introducing a linear contribution to the expected reward:

$$\Delta u_{C,v}\left(k, \zeta_{\text{FOMO}}\right) = \zeta_{\text{FOMO}}\Delta u_{C,v}\left(k\right) + \left(1 - \zeta_{\text{FOMO}}\right)\mathbb{E}\left(V\left(k\right)\right) \tag{21}$$

For tractability, in this TWON prototype we will only include scroll-momentum, setting $\zeta = 1$.

A precise description of this phenomena would require a precise description of the dopaminergic response of each user, and therefore to track the complex, time dependent dopamine level, and how that respond to engagement, novelty and expectations. This complete description goes beyond the scope of TWON. We decided to capture many of the features stemming from scroll momentum by assuming that users can have an extra reward that can increase with the number of content watched consecutively.

The expected net reward, or utility, for consuming a single piece of content k, can be written as the weighted sum of three net rewards:

$$\text{EU}_C\left(k\right) = w_{C,V}\Delta u_{C,V}\left(k\right) + w_{C,R}\left(\left(N_R + 1\right)^{\xi} - N_R^{\xi}\right) + w_{C,t}\Delta u_{c,t}\left(\tau_C\right) \tag{22}$$

where the $w_{C,R}\left(\left(N_R + 1\right)^{\xi} - N_R^{\xi}\right)$ capture the reinforcement mechanism. Values of $\xi \in (0, 1]$ leads to a diminished increase of the *infinite scrolling effect*, leading in most of the case to a spontaneous halting of content consumption. $\xi > 1$ leads to a supralinear reward from *infinite scrolling* leading to users to keep consuming content until a user depleted their budget almost completely [See figure]. $u_{C,V}$ represent the amount of personal value that a user is expecting from consuming a single piece of content. This can be calculated in an analogous way as Eq. 8:

$$\mathbb{E}\left(k\right) = \frac{\sum_l \widetilde{W}_l V_l}{\sum_l \widetilde{W}_l}, \tag{23}$$

where V_l is the value generated by the piece of content k.

4.4.3 Interest and agreement

To quantify the reaction of a user to a piece of content, we need to introduce two main metrics: how interested is the user in the topic, and how much the user likes, or dislikes what has been written in such a topic. Both of these factors can be estimated using properly prompted LLMs, asked to estimate in a scale from 0 to 1 how interested an agent who spoke about certain topics in the past is in the topic of a post, and on a scale ranging from -1 to 1 how much that user likes the specific topic.

Summarizing we have:

$$\mathbb{I}\left(k, \text{past}\right) = f_{LLM,I}(k) \tag{24}$$

$$\mathbb{L}\left(k, \text{past}\right) = f_{LLM,L}(k) \tag{25}$$

In case of a simpler formal model where the opinion on a topic is a real number between 0 and 1, a user would like the topic of post k according to the distance between post k and its own opinion, and

(Here I assume that $\mathbb{I}$ is rescaled between 0 and 1, and $\mathbb{L}$ between -1 and 1) For future developement, we are considering training a ML to automatize the liking/disliking assessment, and tuning the cosine similarity to quantify the closeness in covered topics to measure interest.

4.4.4 Personal value update

Once a user decides to consume a piece of content, its internal counter of the number of consecutive content consumed will increase by one, after this, it will evaluate the actual engagement generated by that piece of content, and it will update its total engagement. For tractability, we decided to quantify the amount of personal value generated by a single piece of content according to how interested the user is in the topic carried by such a piece of content, and by how much the user liked the such a piece of content:

$$\delta E_k = \mathbb{I}\left(k, \text{past}\right) \mathbb{L}\left(k, \text{past}\right), \tag{26}$$

where $\mathbb{I}\left(k, \text{past}\right)$, and $\mathbb{L}\left(k, \text{past}\right)$ are defined in Eq. 24 and Eq. 25 respectively.

But, while the amount of personal value grows linearly with the amount of content consumed, the reward, or utility that the user experience follows the same diminished return discussed in Eq. **??**

4.4.5 Liking/disliking a post or a comment

The decision of liking, or disliking, a post solely depends on the metric $\mathbb{L}$ defined in Eq. 25. Using a simple logistic function without offset wouldn't work, because in average would lead to a user either liking, or disliking every post. Including a "liking threshold" θ could address this problem:

$$P_L(k) = \mathcal{L}\left(\mathbb{L}(k, \text{past}), \beta_L, \theta\right), \tag{27}$$

and

$$P_D(k) = \mathcal{L}\left(\mathbb{L}(k, \text{past}), \beta_D, -\theta\right), \tag{28}$$

For this TWON prototype we assume $\theta = 0.5$, and $\beta_L = \beta_S = 1$. (?values to be checked) Since the time needed for liking or disliking a post is usually negligible, we neglect that contribution for the decision. Furthermore, for the same reason liking, or disliking, a post doesn't interrupt the

4.4.6 Forwarding/sharing a post or a comment

For tractability we will assume that a user will forward, or retweet, a post only if it likes the content of the post, and it's highly engaged in the topic of the post. As mentioned above forwarding a post is a *public* action, therefore is not only subject to the personal preferences of the user in terms of interest and liking, but also on the expected social feedback the user might receive from sharing that piece of content. The net amount of personal reward from sharing a piece of content can be considered proportional to the amount of entertainment that consuming such content generated $u_{S,E} = w_S \delta E_k$. sharing a piece of content requires very little time, so the net time cost for such an action is usually very small:

$$u_{S,t} = \frac{-\tau_S}{b_{i,t}\left(\tau_S - b_{i,t}\right)} \sim \frac{-\tau_S}{b_{i,t}^2} \tag{29}$$

And, finally the reward for social feedback can be written using the same diminished return explained in Eq. 10, where $\mathbb{E}(F)$ is the expected feedback the user is expecting to receive based on the past feedback it received in previous similar posts, forwards, and comments on the same topic, or in general when one doesn't distinguish between topics.

The decision of sharing something will therefore be taken according to the myopic best response model:

$$P_S = \mathcal{L}\left(w_t u_{S,t} + w_f u_{S,F} + w_E u_{S,E}, \beta_S, \theta_S\right), \tag{30}$$

as discussed above, where $\beta_S = 1.5$, and $\theta_S = 0$. (?? Numbers to be double checked).

An important psychological mechanism that has not been included in this framework is the feeling of urgency. If a user is presented with a piece of content that triggers the feeling of urgency it can be shared before analyzing its veracity, leading to the spread of misinformation. And that can lead to dire consequences such as the famous ”Indian Whatsapp lynching”, where people were identified as robbers and beaten to death as a consequence of misinformation content spreading on Whatsapp chats Van der Linden (2023).

4.4.7 Commenting a post or a comment

The decision of commenting on a post is very similar to the one of sharing the post, with two main differences. First, the time needed to write a post or a comment is much larger than the time needed to share such post $\tau_A \simeq \tau_C \gg \tau_S$. Second, the personal value for writing a comment depends on how interest a person is on a topic, and, since a comment could be done both to support an argument or to attack it, increase both for low and for high values of $\mathbb{L}$.

$$E_k = \mathbb{I} \times \mathbb{L}^2 \tag{31}$$

For the sake of tractability, we will consider this personal value at the same level of personal entertainment, and therefore, the reward stemming from this factor is obtained via 12.

4.4.8 Waiting

The last action available to the user is to wait and do nothing for a time τ_W. While this might seem counterintuitive in the frame of maximizing the the entertainment from social media such as youtube, or instagram, it is much more common in social media such as WattsApp or Telegram, where the utility from waiting is to anticipate as much as possible the utility from consuming a piece of content.

4.5 Content creation with LLMs

When an active agent decided to generate content, either by starting a novel conversation, or by responding to earlier contributions by other agents, text is generated by a Large Language Model (LLM). That is, we prompt an LLM with information about the user and, if applicable, the content the agent is responding to.

LLMs not only flooded the consumer market (Teubner et al., 2023) but also academia with text as a research subject (Tiunova and Muñoz, 2023). The abilities of these systems range from classifying and extracting information from unstructured inputs (Xu et al., 2023) to unrestricted text generation adapted to different styles (Bhandarkar et al., 2024). Contemporary research in the social sciences aims to utilize the capabilities to generate content tailored to individual user behavior. A common and predominant approach is to provide an abstract textual description of a political ideology (Argyle et al., 2023). It relies on the model's ability to generalize from abstract ideology description to the appropriate response for generative tasks like social media post generation. The deployment of LLMs as substitutes for humans appears particularly convenient for online social networks (OSNs), as researchers can design an environment that is task-specific and centered around text (Argyle et al., 2023). Thus, measuring polarization tendencies on a large scale with a reproducible approach seems possible. In the current landscape, where OSN providers restrict access to data and obstruct researchers from conducting data-driven experiments based on real data (Bruns, 2021), the synthetic approach may pose a promising solution.

LLMs as synthetic characters The use of LLMs as human simulacra (representation) began with the application as non-player characters (NPCs) in a Sims-style[1] game world to simulate the interpersonal communication and day to day lives (Park et al., 2023). The results showed an authentic but superficially believable human behavior. The current research interest revolves around improving those agents in a technical sense, by refining prompt schemes and model-internal feedback loops (Wang et al., 2024). However, the application of LLMs as synthetic characters has expanded beyond gaming environments into various fields of social science research (Argyle et al., 2023). Researchers are increasingly exploring the potential of these models to simulate human participants in studies, particularly in contexts where obtaining real-world data is challenging or ethically complex. Those disciplines already started to use these models as a replacement in social studies arguing that conditioning through prompting causes the systems to accurately emulate response distributions from a variety of human subgroups Argyle et al. (2023).

Aligning Agents with Human Behavior

Data-driven Modeling UvA provided two datasets that include German Twitter data on political discourses. One dataset consists of Tweets (posts) from delegates of the national parliament concerning political decisions predominantly about the energy transition and the rise of right-wing ideology. The second dataset contains reactions (replies) of regular Twitter users towards these decisions or opinions. Given these two types of data (posts and replies), we derive two tasks and datasets that are learnable for a language model.

[1] The Sims is a series of life simulation video games developed by Maxis and published by Electronic Arts

Composing a post The model is optimized to generate a Tweet based on a given list of topics (words or short statements) and an ideology. The knowledge of the model is bound by the delegate Tweets seen during training. Thus, it is restricted to producing politician-like statements and does not represent the variety of Twitter users.

Commenting on a post The model is optimized to reply given a Tweet (full text) and an ideology. The knowledge of the model is bound by the replies seen during training that predominantly include accusations *(Sie wissen nicht was zielführend ist., Sie haben schon mal so einen Mist geschrieben., Das ist eine dreiste Lüge.)* and insults (*Ihre Platte hat einen Sprung!, Dummheit tötet!, Du bist und bleibst ein dummschwätzender Träumer!)*.

Limitations Our methodological approach reveals important limitations regarding both data selection and the system's capacity to generate discourses containing well-structured standpoints and arguments. While our models demonstrate the capabilities, their performance characteristics primarily reflect the behavioral patterns of the most active users within our dataset, thus mirroring the distinctive communication dynamics observed in the selected Twitter community. This sampling bias raises fundamental questions about the generalizability of our findings across different social media contexts and user populations. The observed patterns prompt several critical considerations regarding discourse quality metrics. To what extent can computational models capture the nuanced variations in argumentation styles across different user communities? The heterogeneous nature of social networks suggests that discourse patterns may vary significantly across different communities, each with its own linguistic norms, interaction styles, and argumentation preferences. This heterogeneity presents both methodological and theoretical challenges. From a methodological perspective, we must consider whether our current modeling approaches are sufficiently sophisticated to capture these variations, or if we need to develop more specialized, community-specific models. Theoretically, we must grapple with the tension between developing generalizable frameworks for understanding online discourse and acknowledging the unique characteristics of distinct social media communities. This raises a broader question about the nature of social media discourse: Are social networks inherently so heterogeneous that meaningful modeling requires a community-by-community approach, rather than attempting to develop universal models of online argumentation?

Symbol	Name	Value	Symbol	Name	Value
i	active user	-	j	targeted user	-
$b_{i,t}$	budget of user i at time t	-	$b_{i,t}^{\text{max}}$	max budget of user i	4h
$b_{i,t}^{r}$	remaining budget of user i at time t within a single round	-	s	resource replenish per round	$b_{i,t}^{\text{max}}\frac{\delta t}{24h} \simeq 0.014$
k_i	post k made by user i	-	s_{k_i}	success of post k	-
s_i	success of user i	-	β_{ON}	slope of the activation function	1.5
$\bar{t}$	time of the last notification	-	$\mathbb{E}\left(T\right)$	Expected time spent online	-
$\mathbb{E}\left(F\right)$	Expected feedback gained if online	-	$\mathbb{E}\left(V\right)$	Expected value gained online	-
$w_{\text{ON},t}$	weight of time cost for going online	1	$w_{\text{ON},f}$	weight of social feedback for going online	1
$w_{\text{ON},v}$	weight of personal value for going online	1	$\bar{t}$	time of last notification	-
τ_G	time for generating a piece of content	10m	τ_C	time for generating a piece of content	5m
τ_R	time for reacting to a piece of content	10s	γ	time constant for memory kernel	1
α	irrationality factor	0	$t^{(l)}$	time of login for the l-th past login	-
T_l	time spent online on the l-th past login	-	$X^{(l)}$	amount of accumulated X in the l-th past login	-
$\widetilde{W}^{(l)}$	weight of the l-th past login	Eq. 7	λ_F	loss aversion for social feedback	1
$w_{f,l}$	weight of the value of the l-th topic for going online	$\delta_{l,0}$	λ_V	loss aversion for personal value	1
$w_{\text{ON,FOMO}}$	weight of FOMO for going online	1	$\Delta\text{EU}_{\text{on}}$	net reward for going online	Eq. 14

Table 3: List of symbols and values for logon. If the value is -, the symbol is a variable, and not a constant

5 The platform model

The platform model comprises all aspects of the TWON that represent design decisions of platform developers. This includes platform affordances such as the presence of like and dislike buttons, the ability to comment or share, or character limits. Furthermore there are algorithms for content ranking, influencing the content each user is exposed to. Additionally, hate speech detection can inform content moderation tools like banning, shadowbanning, or human content moderation can be simulated by limiting the fraction of reviewed content.

In the social sciences, aspects of the platform model would be referred to as ”institutions”, ”the rules of the game in a society or, more formally, the humanly devised constraints that shape human

Symbol	Name	Value	Symbol	Name	Value
EU_G	reward for generating content	Eq. 18	$w_{G,f}$	weight of feedback for generating content	1
$w_{G,t}$	weight of time for generating content	1	β_G	slope of generation function	1.5
$w_{C,R}$	weight of reinforcement for consuming content	1	N_R	consecutive content consumed	$-$
ξ	reinforcement slope coefficient	0.1	$-$	$-$	$-$
$-$	$-$	$-$	$-$	$-$	$-$
$-$	$-$	$-$	$-$	$-$	$-$
$-$	$-$	$-$	$-$	$-$	$-$

Table 4: List of symbols and values for online actions. If the value is -, the symbols is a variable, and not a constant

interaction" (North, 1990). That is, the central decision makers of the TWON are the simulated users. However, platform design has an impact on:

- What activities are possible (e.g. Is there a dislike-button or not?)
- What activities are more or less attractive or costly for a user?
- How likely are users interacting with a specific piece of content?

5.1 Ranking of incoming content

So far, the development of the TWON platform model has focused on the implementation of competing ranking algorithms, as these design aspects are key in the public and scholarly debate about online social networks (Keijzer and Mäs, 2022).

The implemented ranking follows a simple logic. When a user agent i is active in a given round t, the content that has reached this user is sorted according to a score s_c. That is, every piece of content c that the user has received (and not consumed before) is assigned a score. The piece of content with the highest score is ranked highest, and so on.

The score s_c of a piece of content c is computed as detailed in Equation 32. In essence, we assume that every piece of content c can be described by a set of observables that platforms can use to inform their ranking algorithms. These observables include attributes of the content (e.g., the topic, the number of times it has been shared by other users), attributes of the source of c (e.g., the number of followers of the source), and attributes of the receiver (e.g., the frequency of past engagement with content similar to c). The degree to which a given observable is considered by a platform's algorithms is described by w_{obs} ($\in [-1, 1]$), a weight assigned to each observable obs. The term $f_{obs}(c, i)$ describes the degree to which the observable is related to the specific piece of content c and user i. For example, if user i has liked content referring to an issue k many times in the past, and if content c refers to this topic k, then this adds to the score.

$$s_c = f(t) \sum_{obs} w_{obs} \cdot f_{obs}(c, i) \tag{32}$$

We included the following observable, so far:

- A dummy observable that always adopt a weight of 1 (this enables pure chronological ordering)
- number of likes c has received so far
- number of dislikes c has received so far
- number of comments c has received so far
- number of likes comments on c have received so far
- number of dislikes comments on c have received so far
- attributes of the sender of c (success of the sender, similarity of the sender and i)
- attributes of the commenter on c (if a famous person commented on c, it is ranked higher)

To compare alternative ranking algorithms, it is not necessary to implement them. Rather, different sets of weights w_{obs} can be assigned to implement the effect of alternative algorithms on users' actual rankings. To generate a chronological ranking, which will often serve as a baseline for instance, all weights are set to zero except the dummy observable.

Some observables are a counter of interactions, others are continuous measures. For those measures that are based on counting user actions, the impact f_i of a post-related observable i is given as the sum of each instance of the observable, discounted by a factor relating to the time that passed since that instance.

$$f(t) = \text{exp}(-\lambda t) \tag{33}$$

6 Outcome measures

The TWON metrics for online debate quality are presented in Table 5. As elaborated in Deliverables D5.1 and D5.2, the different indicators are non-compensatory: a surplus in one indicator does not automatically make up for a lack in another. Therefore, no single debate quality score will be proposed. Rather, different research projects using this metric can emphasize different aspects/indicators of online debate quality in line with their specific goals. Table 5 presents the metrics in a general form. Specific research projects might be interested in different levels of debate quality. As explained in Deliverable D5.2 these metrics can be amended for use at the individual level (what is the quality of debate observed per individual social media platform user), thread level (what is the quality of a particular thread of comments, debate/topic level (what is the quality of all comments connected to a specific debate, for example, the war in Ukraine) or platform level (what is the quality of debate on the platform as a whole, for example to compare this quality to alternative platforms or platform mechanics). The choice of classification model is based on the evaluation presented in Deliverable D5.2 and explained in Deliverable D5.2.

Table 5: TWON Core Debate Quality Metrics.

Indicator	**Operationalization**
Exposure to political content	Share of comments classified as political with Llama3.1:70b present in the thread to which the participant is exposed
Engagement with political content	Number of political comments liked or shared per participant as classified political with Llama3.1:70b
Contributing political content	Number of comments posted per participant which are subsequently classified as political with Llama3.1:70b
Diversity of exposure	The ideological balance between left, neutral and right-leaning political comments to which a participant is exposed as classified with Llama3.1:70b, if a post is classified as belonging to the minority ideology in a thread it adds a score of 2 to diversity, in the case of a tie it adds 1, otherwise, it adds zero to the cumulative diversity indicator per thread
Quality of exposure	The share of comments to which a participant is exposed which are classified with Llama3.1:70b as substantiating, or expanding on, any claims made within the comment

In addition to the core metrics in Table 5, we also include a set of metrics relevant to debate quality listed in table 6. Like the core metrics, these metrics are evaluated in chapter **??** and explained in chapter **??**.

Table 6: TWON Supporting Debate Quality Metrics.

Indicator	Operationalization
Incivility	Share of comments classified as uncivil with Llama3.1:70b to which a participant is exposed
Interactivity	Share of comments classified as interactive with Llama3.1:70b to which a participant is exposed
Novelty	Share of comments classified as having a new topic compared to previous topics to which a participant is exposed with cardiffnlp/tweet-topic-21-multi

We do not consider these metrics the final answer to measuring debate quality on online social media platforms but a good first step. We are already exploring how to improve them. In Deliverable D5.2, we present findings of ongoing research that show potential to improve the conceptualization and operationalization of the diversity metric.

7 Technical infrastructure

7.1 System overview

The TWON pipeline forms the backbone of this system, facilitating seamless interaction between various components, including the database, the back-end, the front-end, and a machine learning module. This architecture ensures efficient data management, real-time processing, and dynamic decision-making capabilities. Figure 7.1 provides a schematic overview of the technical implementation of the TWON.

At the core of the system lies a MongoDB database, which serves as a secure and scalable storage solution. The database holds critical information such as user interactions, content rankings, and synthetic data generated by agents. Given its flexible document-based structure, MongoDB allows efficient querying and retrieval of large datasets, making it ideal for managing dynamic and evolving user data.

The backend is built using Typescript, ensuring type safety and robust code structure. It acts as a bridge between the front-end and the database, handling API requests, processing business logic, and managing real-time interactions between users and system components. The backend also integrates the machine learning module, which plays a crucial role in agent generation and content ranking.

On the client side, the front-end, also developed in Typescript, provides an interactive and user-friendly interface. It communicates with the backend via well-defined APIs, displaying real-time data,

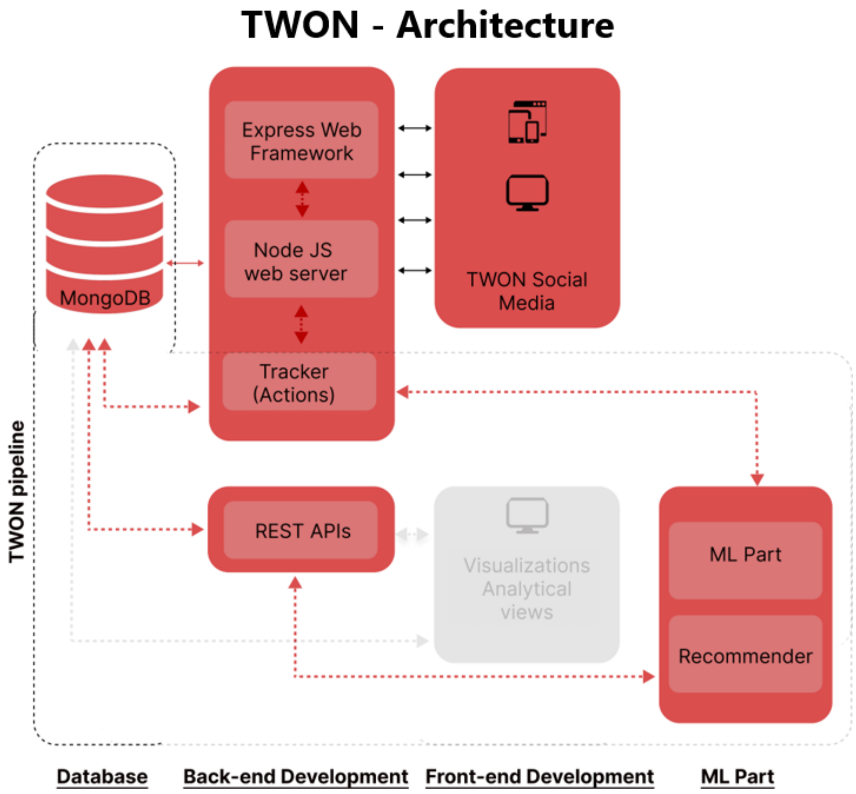

Figure 5: Overview over the System Architecture: Integration of Database, Backend, Frontend, and Machine Learning Module

visualizing system analytics, and enabling users to interact seamlessly with the platform. Typescript improves maintainability and scalability, making the front-end more reliable and efficient.

Additionally, the machine learning module incorporates agent-trained models based on Llama-3.2-3B-Instruct (Llama, 2025). These models are publicly available on the Hugging Face Hub, ensuring seamless integration with the Transformers and PEFT Python libraries. Both models undergo training using the Supervised Fine-Tuning (SFT) approach, which tailors a language model to specific tasks by leveraging labeled data.

The models are optimized for two primary tasks: posting and replying, using content from original users as labeled data. The SFT paradigm refines the model by minimizing token-level errors between the original and generated content, making it an effective technique for enhancing task-specific performance. The training process results in two specialized models:

1. Post Model: Designed to generate user messages (such as tweets) based on a provided list of topics (keywords or short phrases) and an ideological stance. The model's knowledge is limited to the delegate messages encountered during training, restricting it to producing politician-like statements rather than reflecting the broader diversity of users.

2. Reply Model: Optimized to generate replies to messages by considering both the full text of a given messages of another user agent and an ideological perspective. The model's responses are shaped by the replies seen during training, which predominantly include accusatory remarks.

This structured training approach ensures that the models align with the behavioral patterns observed in the original dataset, allowing them to generate contextually relevant content within the defined ideological constraints.

Overall, this architecture ensures that all system components work together cohesively, with MongoDB providing reliable data storage, and Typescript-based back-end and front-end enabling smooth interactions. This integrated approach allows for a highly scalable, efficient and intelligent system capable of handling complex user interactions and content ranking in real-time.

7.2 Simulation services

The system comprises three core services that work together to simulate user interactions and content ranking. Network Generation Service, Agent Scheduler, and Ranking Service. These services collectively manage the lifecycle of agents (users), the structure of user networks, and the ranking of content before any user action is performed. The generated data are stored in the MongoDB database (see Figure 7.2).

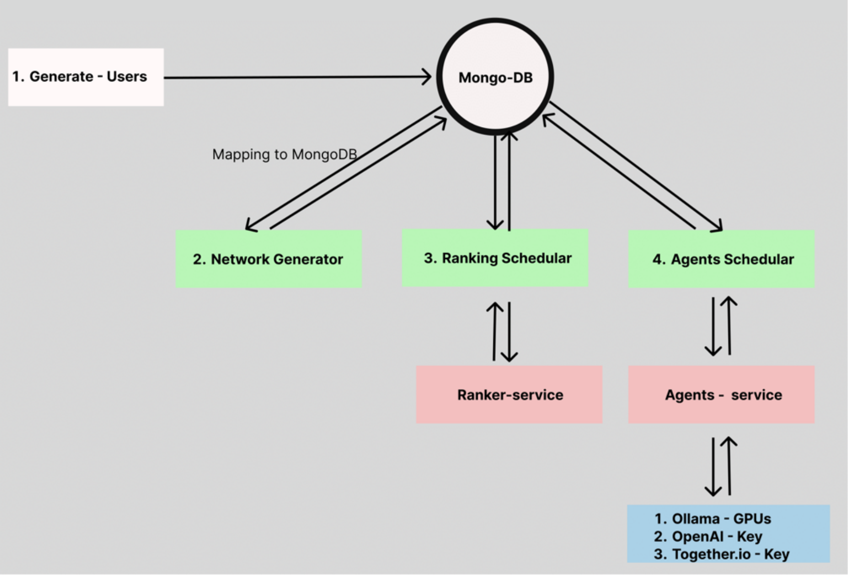

Figure 6: Integrated Workflow of Network Generation, Ranking Scheduler, and Agent Scheduler Services with Secure Data Storage

7.2.1 Network generators

The Network Generation Service is responsible for creating and maintaining the underlying user network and, thus, determines who may interact with whom. Whether two agents connected by a network link actually do interact (e.g. communicate messages) depends on their own decisions and decision of the platform (e.g. ranking of messages).

At the moment, we have implemented the most standard network generators, including the Erdős–Rényi and Barabási–Albert algorithms (Erdos et al., 1960; Barabási and Albert, 1999). The Erdős–Rényi model generates random connections between users, while the Barabási–Albert model forms scale-free networks, where a few highly connected nodes (hubs) emerge over time. These network structures play a crucial role in determining how information flows between users and how interactions are influenced within the system.

7.2.2 Ranking

When an agent is activated, the TWON determines what content the agent is exposed to and the ordering/ranking of this content. Computationally, this is implemented in the Ranking Service (see also Section 5). This service ensures that content is ranked based on specific criteria, such as relevance, popularity, or engagement levels. Before each action an agent takes, the system dynamically updates rankings, allowing the simulation to reflect real-world content visibility mechanisms.

7.2.3 Agents Scheduler

The Agent Scheduler governs the activities of virtual agents within the system. That is, it manages the user model (see Section 4) of the TWON, including behavioral aspects such as login probability, motivation levels, and time budget, which determine when and how agents engage with the platform. Once logged in, agents can perform various actions, including posting content, commenting on existing posts, liking, or disliking posts. Each action is influenced by predefined behavioral parameters, ensuring a realistic simulation of user interactions over time.

7.3 Infrastructure and Deployment

TWON is built using Typescript and is designed for large-scale social media simulations powered by LLM-driven agents. The platform is deployed as a scalable application, enabling researchers and developers to analyze agent behaviors, interaction patterns, and emergent dynamics in a controlled environment. It ensures efficient agent lifecycle management, monitoring agent activities, resource consumption, and interaction states.

For seamless deployment, TWON leverages Podman, a container management tool that provides lightweight, rootless, and secure execution environments, ensuring greater flexibility and isolation (Podman, 2025). This setup allows TWON to be easily deployed, scaled, and managed across different environments while maintaining security best practices. Additionally, TWON features real-time analytics, tracking key metrics such as agent activity duration, and post frequencies. Built on Node.js (v16 or higher) with TypeScript 4.x, TWON provides a scalable and adaptable architecture for studying complex social media interactions (see the implementations on Github).

7.3.1 Medium scale simulations

We conducted our initial set of experiments using this simulation framework to analyze various aspects of agent behavior and system dynamics. Simulation was conducted with 5 agents on the topic of elections in Germany. LLM-agents were fitted to the Germany's Twitter data on politicians and we simulated the "effect of Musk" on Twitter. These preliminary experiments allowed us to observe how agents interact within the simulated environment, providing insights into their lifecycle, time allocation, and motivational factors influencing their actions. By running controlled simulations, we captured key behavioral trends and emerging patterns, offering a comprehensive overview of how agents engage with content, react to different stimuli, and participate in social media interactions over time.

To further illustrate these findings, we present a series of graphs that visualize critical aspects of the simulation. These include an analysis of agent lifecycles, showing the duration and transitions between active and inactive states (see Figure 7.3.1), as well as a breakdown of agents' time budgets (see Figure 7.3.1), highlighting how they allocate their available time for different tasks. Additionally, we examine motivation score (see Figure 7.3.1), capturing the factors influencing agent decision-making, alongside a detailed timeline of agent actions such as posting, commenting, liking, and disliking (see Figure 7.3.1). Furthermore, we present ranking frequency metrics (see Figure 7.3.1), which demonstrate how often content ranking updates occur before agent actions, and user interaction frequency graphs, providing insights into the engagement dynamics within the simulation (see Figure 7.3.1). These visualizations help to better understand the underlying patterns governing agent behavior and the broader system mechanics.

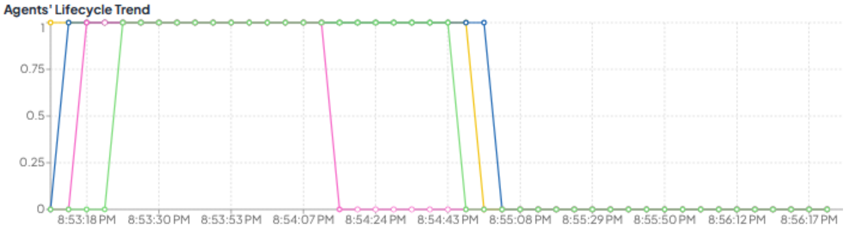

Figure 7: Agents' Lifecycle: Duration of Activity and Inactivity Phases. An example with five agents

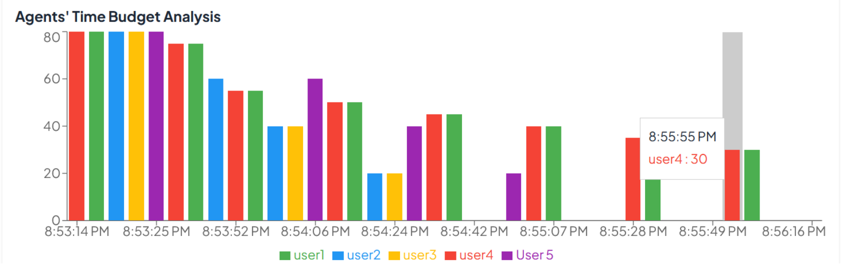

Figure 8: Agents' Time Budget: How the time budget is increasing/decreasing. An example with five agents.

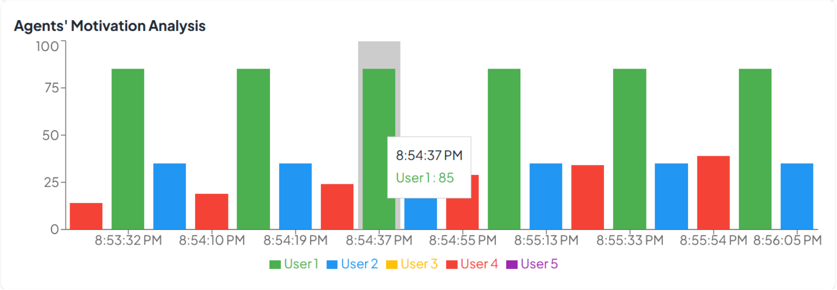

Figure 9: Agents' Motivation: How the motivation is increasing/decreasing. An example with five agents.

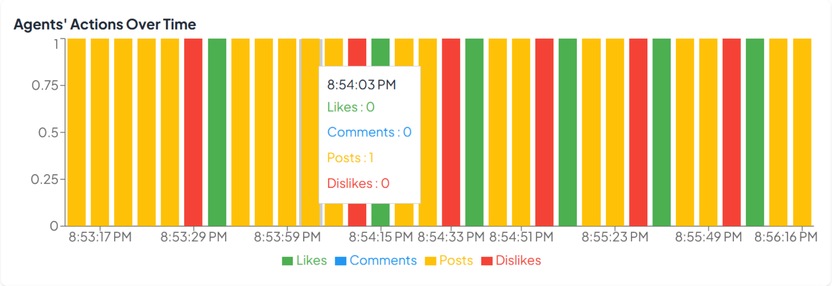

Figure 10: Agents' Actions: actions (post, comment, like, dislike) by five agents over time. An example with five agents.

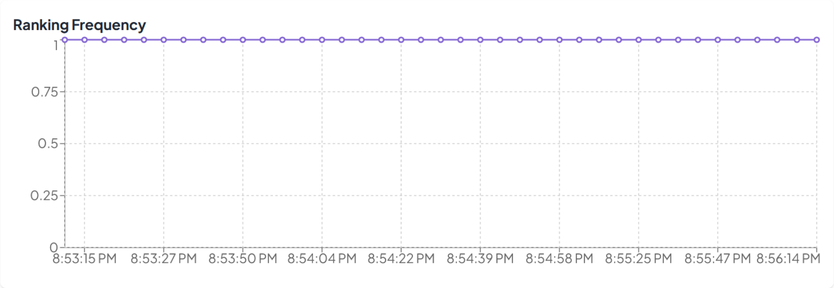

Figure 11: Content Ranking Frequency: This figure illustrates how often the content is ranked.

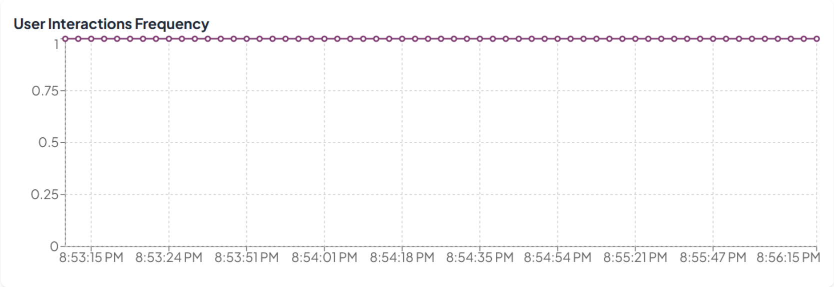

Figure 12: "User History Collection Before Action: This figure illustrates the frequency at which user history was generated."

References

Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3): 337–351, 2023. 39

Robert Axelrod. The dissemination of culture: A model with local convergence and global polarization. *Journal of conflict resolution*, 41(2):203–226, 1997. 12, 13

Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439): 509–512, 1999. 48

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. 7

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA, 2019. Association for Computational Linguistics. doi: 10.18653/v1/S19-2007. URL `https://www.aclweb.org/anthology/S19-2007`. 18

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. 8

Michael Belfrage, Emil Johansson, Fabian Lorig, and Paul Davidsson. [in] credible models–verification, validation & accreditation of agent-based models to support policy-making. *JASSS: Journal of Artificial Societies and Social Simulation*, 27(4), 2024. 10

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. 8

Avanti Bhandarkar, Ronald Wilson, Anushka Swarup, and Damon Woodard. Emulating author style: A feasibility study of prompt-enabled text stylization with off-the-shelf llms. In *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*, pages 76–82, 2024. 39

Axel Bruns. *Are filter bubbles real?* John Wiley & Sons, 2019. 7

Axel Bruns. After the ‘apicalypse’: Social media platforms and their fight against critical scholarly research. *Disinformation and Data Lockdown on Social Platforms*, pages 14–36, 2021. 39

Joshua M Epstein. Why model? *Journal of artificial societies and social simulation*, 11(4):12, 2008. 10

Paul Erdos, Alfréd Rényi, et al. On the evolution of random graphs. *Publ. math. inst. hung. acad. sci*, 5 (1):17–60, 1960. 48

Ernst Fehr and Klaus M Schmidt. A theory of fairness, competition, and cooperation. *The quarterly journal of economics*, 114(3):817–868, 1999. 22

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. 8, 12

Douglas D Heckathorn. The dynamics and dilemmas of collective action. *American sociological review*, pages 250–277, 1996. 23

Sophia Horn, Sven Banisch, Veronika Batzdorfer, Andreas Reitenbach, Fabio Sartori, Daniel Schwabe, and Michael Mäs. Success-driven user activity contributes to online polarization. *Preprint available at SSRN: https://ssrn.com/abstract=5031685 or http://dx.doi.org/10.2139/ssrn.5031685*, 2024. 13, 14, 16

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`. 7

Marijn A Keijzer and Michael Mäs. The strength of weak bots. *Online Social Networks and Media*, 21: 100106, 2021. 12

Marijn A Keijzer and Michael Mäs. The complex link between filter bubbles and opinion polarization. *Data Science*, 5(2):139–166, 2022. 6, 7, 42

Marijn A Keijzer, Michael Mäs, and Andreas Flache. Communication in online social networks fosters cultural isolation. *Complexity*, 2018(1):9502872, 2018. 11

Konstantin Klemm, Víctor M Eguíluz, Raúl Toral, and Maxi San Miguel. Global culture: A noise-induced transition in finite systems. *Physical Review E*, 67(4):045101, 2003. 15, 16

Siegwart Lindenberg. The paradox of privatization in consumption. In *Paradoxical effects of social behavior: essays in honor of anatol rapoport*, pages 297–310. Springer, 1986. 22

Meta Llama. Llama 3.2-3b instruct, 2025. URL `https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct`. Accessed: 2025-02-25. 47

Michael Mäs. Analytical sociology and complexity research. In *Research Handbook on Analytical Sociology*, pages 100–118. Edward Elgar Publishing, 2021. 12

Michael Mäs and Dirk Helbing. Random deviations improve micro–macro predictions: An empirical test. *Sociological methods & research*, 49(2):387–417, 2020. 24

Michael Mäs and Heinrich H Nax. A behavioral study of "noise" in coordination games. *Journal of Economic Theory*, 162:195–208, 2016. 24, 26

Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444, 2001. 14

Robert K Merton and Paul F Lazarsfeld. Friendship as a social process. *M. Berger (sous la direction de), Freedom and control in modern society. New York, Van Norstrand*, 1954. 14

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, pages 1–17, 2018. 18

James Moody and Douglas R White. Structural cohesion and embeddedness: A hierarchical concept of social groups. *American sociological review*, 68(1):103–127, 2003. 18

Douglass C. North. *Institutions, Institutional Change and Economic Performance*. Cambridge University Press, Cambridge; New York, 1990. 42

Barack Obama. Farewell adress, 2017. URL `https://obamawhitehouse.archives.gov/farewell`. 6

Karl-Dieter Opp. Contending conceptions of the theory of rational action. *Journal of theoretical politics*, 11(2):171–202, 1999. 22

Johan Ormel, Siegwart Lindenberg, Nardi Steverink, and Lois M Verbrugge. Subjective well-being and social production functions. *Social indicators research*, 46:61–90, 1999. 22

Eli Pariser. *The filter bubble: What the Internet is hiding from you*. Penguin UK, 2011. 6

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22, 2023. 39

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. 7

Podman. Podman: A tool for managing containers, 2025. URL `https://podman.io`. Accessed: 2025-02-25. 49

Adil Rasheed, Omer San, and Trond Kvamsdal. Digital twin: Values, challenges and enablers from a modeling perspective. *IEEE access*, 8:21980–22012, 2020. 7

Fabián Riquelme and Pablo González-Cantergiani. Measuring user influence on twitter: A survey. *Information processing & management*, 52(5):949–975, 2016. 13

Sara Rosenthal, Noura Farra, and Preslav Nakov. Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, pages 502–518, 2017. 18

Herbert A Simon. Bounded rationality. *Utility and probability*, pages 15–18, 1990. 23

Franz-Walter Steinmeier. Christmas message, 2017. URL `https://www.bundespraesident.de/SharedDocs/Reden/EN/FrankWalter-Steinmeier/Reden/2018/12/181225-Christmas-message.html`. 6

Timm Teubner, Christoph M Flath, Christof Weinhardt, Wil van der Aalst, and Oliver Hinz. Welcome to the era of chatgpt et al. the prospects of large language models. *Business & Information Systems Engineering*, 65(2):95–101, 2023. 38

Alla Tiunova and Felipe Muñoz. Chatgpt: Using ai in social studies academic research. *Available at SSRN 4451612*, 2023. 38

Sander Van der Linden. *Foolproof: Why misinformation infects our minds and how to build immunity*. WW Norton & Company, 2023. 38

Cynthia Van Hee, Els Lefever, and Véronique Hoste. Semeval-2018 task 3: Irony detection in english tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 39–50, 2018. 18

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345, 2024. 39

Neil D Weinstein. Unrealistic optimism about future life events. *Journal of personality and social psychology*, 39(5):806, 1980. 27

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. 7

Fang Wu, Dennis M Wilkinson, and Bernardo A Huberman. Feedback loops of attention in peer production. In *2009 International Conference on Computational Science and Engineering*, volume 4, pages 409–415. IEEE, 2009. 13

Shicheng Xu, Liang Pang, Huawei Shen, and Xueqi Cheng. Nir-prompt: A multi-task generalized neural information retrieval training framework. *ACM Transactions on Information Systems*, 42(2):1–32, 2023. 38

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, 2019. 18

Ekaterina Zhuravskaya, Maria Petrova, and Ruben Enikolopov. Political effects of the internet and social media. *Annual review of economics*, 12:415–438, 2020. 7