New Automation for Social Bots: From Trivial Behavior to AI-Powered Communication

Christian Grimme[1(✉)], Janina Pohl[1], Stefano Cresci[2], Ralf Lüling[3], and Mike Preuss[4]

[1] University of Münster, Münster, Germany
{christian.grimme,janina.pohl}@uni-muenster.de
[2] IIT-CNR, Pisa, Italy
s.cresci@iit.cnr.it
[3] Aleph Alpha, Heidelberg, Germany
ralf.lueling@aleph-alpha.de
[4] LIACS, Universiteit Leiden, Leiden, The Netherlands
mpreuss@liacs.leidenuniv.nl

Abstract. Today, implications of automation in social media, specifically whether social bots can be used to manipulate people's thoughts and behaviors are discussed. Some believe that social bots are simple tools that amplify human-created content, while others claim that social bots do not exist at all and that the research surrounding them is a conspiracy theory. This paper discusses the potential of automation in online media and the challenges that may arise as technological advances continue. The authors believe that automation in social media exists, but acknowledge that there is room for improvement in current scientific methodology for investigating this phenomenon. They focus on the evolution of social bots, the state-of-the-art content generation technologies, and the perspective of content generation in games. They provide a background discussion on the human perception of content in computer-mediated communication and describe a new automation level, from which they derive interdisciplinary challenges.

Keywords: Social media · Automation · Bots · Artificial intelligence · Content generation

1 Introduction

Automation[1] in social media has become a central point of discussion in computational social science, computer science, psychology, political science, journalism, and related domains. The central questions are *whether*, *to what extent*, *how*

[1] In the most general sense, we understand automation to mean technically controlled processes that ensure a specified target achievement largely without human intervention. In closed-loop systems, target achievement is controlled by feedback mechanisms and through self-regulating control mechanisms. In open-loop systems, no feedback mechanism is implemented [31].

F. Spezzano et al. (Eds.): MISDOOM 2022, LNCS 13545, pp. 79–99, 2022.
https://doi.org/10.1007/978-3-031-18253-2_6

convincingly, and *with what effects* automation is used. In the online and social media ecosystem, automation usually relates to the (partly or completely) self-regulating mechanization of communication by algorithms, either with a wider public (one-to-many) or individuals (one-to-one). Today, an almost classic subject of discussion is the social bot – a type of automaton in online media that (also temporarily) hides behind (possibly real) online accounts to interfere with human communication [14,30]. Those bots are often described as tools that act autonomously, behave or respond intelligently to others, and even manipulate people's minds, e.g., to influence elections [8]. Others describe social bots and automation as simple tools, performing simple tasks like amplifying predefined content designed by humans [6]. Still others narrow their view to the assumed intelligent and autonomous automata, find that social bots do not exist at all, and claim that contemporary research on this topic is a huge conspiracy theory by greedy scientists aiming for funding [27]. Regardless of how social bots are defined and understood (as simple spammers or as super-smart and autonomous actors), we ask the question whether automation is already mature enough to produce textual and pictorial content systematically, autonomously, and convincingly[2] so that people can be manipulated by their means?

This work addresses the context of automation in online media with a broad perspective. It refrains from finding definitions of social bots, deficiencies in methodology, or participating in the non-targeted discussion of conspiracy beliefs. On the contrary, it focuses on exploring the potential of automation in the online media ecosystem based on current technologies and preconditions and discusses potential threats and challenges that may arise as these technologies are combined or advance further. Clearly, the authors are convinced of the existence of automation in social media. However, they acknowledge (and have contributed to) the discussion on shortcomings of current scientific methodology in investigating this phenomenon (e.g. in detection methods, [7,29,30]).

The presented results and discussion partly stem from a theme development workshop titled "AI: Mitigating Bias and Disinformation", held in May 2022, which also addressed the topic of "Automation in Online Media". The authors of this paper are a subset of experts and participants of the workshop, approaching communication automation from different angles by integrating the evolution of social bots, state-of-the-art content generation technologies, and the perspective of content generation in games. In the remainder of this work, we provide a background discussion on the human perception of content in computer-mediated communication. After that, we describe a new automation level and derive interdisciplinary challenges from it. Overall, we present this paper as an initial perspective on mid-term and future challenges and research questions regarding possible new aspects of automation.

[2] "Convincingly" in the sense that social media users are not aware of messaging with an automaton or consuming artificially generated content. This does not relate to direct change of opinion.

2 Background and Context of Computer-Mediated Communication

Although the context of automation in online media is about communication, the infrastructural and technical conditions of social media platforms provide a unique and, in contrast to direct human communication, often limited form of communication. Consequently, communication behavior, content, and reception deviate, influencing the requirements for the automation of communication. Over the years, several theoretical frameworks have been developed that may illuminate aspects of human-computer communication and interaction patterns.

A basic model of computer-mediated communication (CMC) developed by Walther [76] states that computer interfaces enable humans to communicate with less bias. Due to the reduced number of cues (e.g., absence of ambient noise or gestures), fewer social responses are triggered that blur the transmitted message. However, recent works show that the number of social media cues increased dramatically, e.g., due to the introduction of emojis [78]. These shall provide rich information, e.g., to judge whether someone else is a fake account [34].

Nass et al. developed the computers are social actors (CASA) model [49]. CASA states that cues emitted by a machine can trigger users to apply the same social heuristics used for human-to-human interactions [59]. Studies prove CASA's validity in the social media context: Ho et al. [33] showed that chatbots and humans could be equally effective in achieving positive conversation outcomes. Similarly, bots posting informational content on Twitter are perceived as trustworthy, although human curators were assessed as being more credible [22,23,66].

The Uncanny Valley, introduced as a theoretical (and partly speculative) model by Mori [47], is defined as the low point of a qualitative function representing human affinity toward technology. Generally, affinity increases the more human-like machines become. This, however, only works until their real nature is unclear to a human observer, which then provokes a feeling of eerie. The brain triggers this effect when incoherent behavior is detected, i.e., when the expectation of seeing a machine is not met by reality [62]. In line with CASA, participants in a study by Skjuve et al. [65] reacted similarly to a bot as to a human conversation partner, as long as bots were able to carry the conversation. In several other studies, people reacted more positively to a chatbot without an avatar or were more likely to befriend another user with a comic-like rather than a hyper-realistic virtual avatar [5,13,63].

Sundar's MAIN model [68] defines cues that are used mindlessly by humans to rate digital media's credibility: Modality, Agency, Interactivity, and Navigability. Intuitively, an audio-visual mode has more credibility than text-only media. However, presumably because of the Uncanny Valley effect, people trust multimodal media, including text and images, although audio-visual media are closer to real conversations. Agency is defined as the source of information, i.e., the more social presence the source has, the more trustworthy it is. Interactivity addresses the response behavior of digital media. The faster and more adapted

the response to the ongoing conversation is, the more trust is granted. Navigability features the design of a digital medium, i.e., information structured according to human expectations is more credible. Using the MAIN model, researchers showed that humans perceive chatbots as being credible [66], while bots on Twitter found with automated detection methods like Botometer [80] are considered less credible [4]. Nevertheless, since Botometer was shown to detect only unsophisticated bots [15,29], the limited cues given by them (e.g., text-only, no variation of actions) may lead to the reduced credibility assessment.

Already before the COVID-19 pandemic and Facebook announced its transformation into Meta Platforms [45], the Metaverse has been considered as digital future of social interaction [39]. The term was coined in a science-fiction novel from 1992 [67] to describe a virtual world next to the physical world in which users interact with each other and services via avatars. It combines elements from virtual and extended reality (VR and XR, respectively) but is not adequately defined yet due to the use of the word for different marketing purposes [57]. First steps towards a Metaverse have been taken in products of the gaming industry such as Second Life or online role play games [40]. Recently, Facebook published its first version of a Metaverse in which humans can create avatars and socialize with friends while wearing VR headsets [53]. Due to the resurgence of the topic, little contemporary research has been published until now. Jeon [36], for example, studies how users designing their perfect self in a Metaverse react to advertisements emotionally, while others explored the security and privacy risks of the Metaverse [20].

The development of theories about how humans perceive the digital world and act in it happened in line with the advancement of technologies. Due to various social cues transmitted via social media but also due to known restrictions and the human ability to bridge perception gaps with social scripts (anthropomorphization), humans may perceive the online ecosystem as similar to the real world, especially if no unexpected behavior occurs.

3 Three Perspectives

To illuminate the current state and future perspectives of (automatic) communication on social media, we discuss three different viewpoints: social bots as actors, content generation models as tools, and games and artificial intelligence as references for content generation in virtual worlds. Especially the interplay and interference of these three perspectives provide a multifaceted basis for identifying current unresolved issues and future challenges.

3.1 Evolution of Social Bots

The paramount example of automation in social media is the social bot – an account that is at least partially automated to perform a set of predefined tasks. Since the very emergence of social media, their support for anonymity and the possibility of setting up programmatic interactions via APIs resulted in the rapid

development and diffusion of social bots [26]. Despite the existence of neutral or even benign bots that contribute to answering the information needs of social media users, a large number of bots have shady purposes. Because of this, and in parallel to the rise of social bots, platform administrators and scholars devoted significant efforts to the development of bot detection techniques [14,44].

Through time, the characteristics of social bots have changed much. Bots developed in the early 2010s were very simple accounts characterized by limited personal information, few social (i.e., friend/follower) relationships, and repetitive posting activity. On the one hand, their simplicity allowed bot developers to create many such accounts in a short time. On the other hand, however, it also made detecting those bots a relatively easy task [79]. For this reason, subsequent social bots were more sophisticated, featured detailed – yet obviously fake – profile information (e.g., credible profile picture, short bio or account description, birthday), and had human-comparable social relationships and diversified activity. These characteristics made the sophisticated social bots much harder to be distinguished from human-operated accounts, as empirically demonstrated by the increased difficulty of both social media users and machine learning-based bot detectors at spotting newer bots with respect to older (and simpler) ones [15]. In fact, the development of sophisticated social bots started an arms race between bot developers and bot detectors that continues these days [16].

The burden of creating carefully engineered and thus credible bots were on the shoulders of the bot developers. In other words, all of the detailed information required to disguise social bots had to be manually inserted, which implies that significant effort and time were required to create a large number of sophisticated bots. Similarly, the behavior of social bots was rule-based, meaning that bot developers typically created simple sets of rules to determine their actions and activities. These could drive the bots to reshare all content posted by certain accounts, post messages at predefined times, or even automatically follow a set of target accounts. Overall, until recently, social bots featured limited "intelligence", independently of their complexity and degree of resemblance to human-operated accounts [6]. However, this scenario is about to change due to the recent advances in AI that provide unprecedented opportunities for creating more intelligent and human-like social bots. For example, generative adversarial networks (GANs) demonstrated exceptional capabilities at artificially creating realistic-looking pictures of men and women of all ages,[3] among other things.[4] These could very well be used as credible profile pictures of fake accounts, as it already happened on Facebook and Instagram.[5] Similarly, recent advances in natural language generation (e.g., OpenAI's GPT 3) opened up the possibility to create artificial texts on any given topic,[6] even mimicking the writing style of a target character, or adopting a peculiar one. New bots could (and already do) exploit these techniques to craft more effective and credible messages before posting them on social media [25,52]. Finally, AI has also been used to generate

[3] https://thispersondoesnotexist.com/.
[4] https://thisxdoesnotexist.com/.
[5] https://www.wired.com/story/facebook-removes-accounts-ai-generated-photos/.
[6] https://openai.com/blog/better-language-models/.

artificial online behaviors (i.e., sequences of actions) to trick detectors of malicious accounts into misclassifying AI-driven accounts as benign ones [32]. These figures paint a worrying picture of the capabilities that future bots could exhibit.

3.2 Multimodal Artificial Content Generation

The advent of transformers-based language models like BERT [18] or GPT [54] changed the status-quo of natural language generation (NLG). In contrast to previous approaches like convolutional neural networks, transformers draw global dependencies between input tokens, allowing the connection of coherent words that do not appear in consecutive order [75]. Additionally, by using as many unlabeled, cross-domain, and multilingual texts as possible during an extensive pre-training, transformers gain a good understanding of language and implicitly learn a variety of potential sub-tasks. Thus, few- or even zero-shot learning is possible, where the model either receives only a few examples as input or even fulfills the task spontaneously [11].

The current state-of-the-art in text-only generation is GPT-3 [11], which can be used to generate texts that are indistinguishable from human-written ones, especially if they are short [35]. The mean human accuracy at detecting five hundred word articles written by GPT-3 was 52% [11][7]. Although a BERT model trained to detect GPT-generated texts performed slightly better, finding a reliable way to detect these artificial texts remains an open task [1]. Fagni et al. [25] demonstrate this problem based on fake accounts that use artificial tweets generated with GPT-2, amongst others. They evaluated thirteen supervised detectors, like various BERT variants, assessing several accounts and tweet features. Accounts backed-up with GPT-2 generated tweets were hardest to detect for these trained models, with a mean accuracy of 75%.

However, in the context of the automated production of information, not only the text is relevant, but also the associated visualizations in the form of images, drawings, and avoidably scientific diagrams to underline the statements to be conveyed. Large language models can help generate natural language treatises to generate the associated visualizations and then describe them depending on the situation. A good way to generate images from text is Dall-E, a 12 billion parameter version of GPT-3 that is trained to generate images from text descriptions using a data set of text-image pairs [55]. Extensive pre-training is fast becoming the norm in Vision Language (VL) modeling. However, the prevailing VL approaches are limited by the need for labeled data and the use of complex multi-level pre-training targets. It is a simple method for enriching generative language models with additional modalities using adaptor-based fine-tuning. For example, building on Frozen [74], the Aleph Alpha model MAGMA [24] trains a set of VL models that autoregressively generate text from any combination

[7] The readers may ask themselves whether they can judge who wrote the abstract of this paper - the authors or GPT-3. In fact, the abstract has been generated automatically by GPT-3 using only the introduction chapter of this paper as input. No editing has been done by the authors.

of visual and textual inputs. The pre-training is fully end-to-end and uses a single language modeling objective, which simplifies optimization compared to previous approaches. Notably, the language model weights remain unchanged during training, allowing for the transfer of encyclopedic knowledge and contextual learning skills from language pre-training.

3.3 Perception of Content in Games and Social Media

Games are at the forefront of AI research and have recently been a testbed for many new algorithmic developments, which have led to seminal papers. Deep Reinforcement Learning was first shown to be successful on the Atari Learning Environment [46], and the more abstract (board game) Go problem was first successfully tackled on and beyond the human grandmaster level using AlphaGo [64]. Many improvements followed, as summarized in [61]. Togelius [71] explains why this direction is going to continue to be prominent in AI, especially if we want our methods to further develop in the direction of *artificial general intelligence* (AGI). As of May 2022, the last current step may be Gato [58], an agent that can deal with hundreds of tasks, including many games successfully, but also handle natural language problems. In this case, solving a problem often means to *create* an answer that matches the expectations of humans. Whereas Gato can e.g. create speech that matches the context as, e.g., GPT-3 [11] does, Dall-E2 [55] generates stunning pictures from text prompts.

Generation of content has some tradition in the computer game field; it has been one of the most vital research areas in this realm at least since around 2013 [72] and is usually subsumed under the term *Procedural Content Generation* (PCG). There are early examples of generative methods for maps/levels in games already employed in the 1980s, notably Elite which featured a vast science fiction universe that could by no means have been stored in the memory of available computers. The generation method basically relied on controlled randomness, however, more recent methods use randomness only as variational effect to prevent too strong similarities in the provided content which would create an "artificial" impression. As a main driving force, they use explicit optimization (according to a measurable criterion) or a model that implicitly stores knowledge about content in a machine learning fashion, usually a (deep) artificial neural network. Nowadays, there is basically no type of game content that is not semi-automatically or fully automatically generated to some extent, including whole non-player characters (NPCs), missions or full plots, graphical components, music up to almost complete game creation. Notable examples here are *No Man's Sky*[8] (2016, as of 2022 still extended several times every year), and Ultima Ratio Regum [38][9] (started in 2012, still in beta). Content creation may also be personalized to the expectations of users according to the Experience-Driven PCG paradigm [81].

From this viewpoint of users, and especially if seen from an automated generation perspective, computer games share a lot with social media:

[8] https://www.nomanssky.com/.
[9] https://www.markrjohnsongames.com/games/ultima-ratio-regum/.

interactivity: whereas most of the content has to be perceived by the user, interaction is not only possible in contrast to other media (e.g., movies, newspapers), but a vital component of the setting;
immersion: games, as well as social media providers, aim to catch and hold the attention of users as long as possible;
believability: it is not necessary to understand the content or use it with a specific plan or intention, but it must be made in a way that appears to be meaningful and believable.

One crucial difference, especially concerning believability, may be that in computer games, users apply the *suspension of disbelief* because they know that they are in a game's context and still *want to believe* in the content they see. In other words, they know that they will be tricked but want to be tricked well enough to ignore that thought. In social media, users usually expect to be confronted with believable content because it is real, produced by other users with some intention. It seems necessary to make quite big mistakes to raise the user's suspicion that the observed content may be generated, which, of course, simplifies betraying users by inserting (semi-) automatically created content and making them believe it is from real users. Thus users do not expect to be tricked, and therefore the level at which small mistakes go unnoticed is relatively high.

In games and social media, being successful requires achieving emotional attachment to provoke reactions. However, and this is another difference, the attachment must be at least partly positive in games. It can be challenging, but players will simply churn and play another game if it is more negative than positive. This is not the same for social media users who are also engaged by negative attachment (e.g., shit storms). Additionally, social media content is consumed at a much higher frequency, user attention is much more fluent, and several threads can be worked on in a minute. Therefore, believability may be more effortless to achieve as the amount of content a user sees before, e.g., accepting or setting up a friendship request or supporting an existing statement, is relatively small.

In consequence, we can presume that making believable content in social media, especially for fast-paced media like Twitter (where lengthy statements are rare), is probably easier than for games, where the problem of computationally generating narrative is still only working in specific contexts and on a small scale [2]. Additionally, considering that some human social media actors (e.g., from the Alt-Right scene) use distortion and confusion as means of communication, it seems even easier to produce believable postings automatically. Generating nonsensical, out-of-context, or arbitrary statements is certainly possible already now with the available generation algorithms, as GPT-3 [11]. Despite these advances in generation, putting different media types together is undoubtedly more challenging. In game AI, this is known as facet orchestration [41] with the overall goal of generating full games, and there are only a few examples of doing it only with two facets (e.g., graphics and audio) successfully, none of which goes into a completely automated direction. A certain amount of human coordination is always necessary to obtain a good result. Using techniques such

as Dall-E2 or MAGMA would not help here, as they would just try to express the same content in another facet (from text to graphics or vice versa), but in games and media, text and pictures used in the text are not totally congruent but rather synergetic.

Riedl already argued in 2016 [60] that being able to generate meaningful computational narrative is necessary for interaction with humans. This leads towards the probably most important current research direction in AI, which deals with the cooperation of AI agents and humans. It comes under different labels, human-computer interaction, hybrid intelligence, team AI [48], computer-supported cooperative work, but eventually means that machines have to interact in a meaningful way with humans and other machines even if an out-of-distribution event (something they have not been trained for) happens. Thorough research in this direction has just started and presumably, will keep us busy for a long time. In the meantime, interaction with the AI that controls a bot will be the only (fairly) safe way for a human user to find out if there is a machine on the other side, as was suggested for games some time ago [43]. Needless to say, Turing-testing is itself hard to automatize, making it a cure for experienced users but not for automated bot-finding.

4 New Automation

In the previous sections of this paper, we presented various developments, research results, models, and insights regarding the current state of automation in online media, the generation of artificial content, and the perception of content and communication in various technical environments (social media to game worlds). At this point, we draw new conclusions from these observations and point out what seems to be a realistic perspective toward a new level of automation. New Automation creates challenges that go far beyond previous research questions and will need to be addressed by the research community and society in the future.

Automation in social media is currently, in most cases, still limited to the technical implementation and imitation of human behavior at a rudimentary level. Besides the massive content duplication, only simple reactive actions are usually performed on other users (repetition of content, signaling approval/disapproval). In this context, the simulation of human-like behavior does not primarily serve to increase the credibility of automated actors vis-á-vis human communication partners but rather to avoid detection and sanctioning by monitoring mechanisms of social platforms. Similarly, massively repetitive content and automated approval or disapproval do not aim at human communication partners. They target the recommendation mechanisms that decide which content and topics users see as important in their timelines [69].

This status quo may now change permanently under the new circumstances of the development of content-generating technologies. While it was previously challenging to generate thematically appropriate content without human intervention, transformers-based neural networks and even multimodal advancements of

these technologies now represent a step in the direction of (partially) autonomous and reactive systems for direct communication. Thus, the technological base of behavioral imitation can now be complemented by a substantive building block of automation in content generation (see Fig. 1). Specifically, automata can now be developed so that they not only behave in a human-like manner (i.e., follow a regular daily routine, simulate human reaction speed) but also generate creative-seeming but indeed variable content. Textual content is not only variable at the word level; it can also be preset to views and opinions to a limited extent (few-shot learning). For this purpose, content from other users can be used as preset content to configure an opinion of the automaton to simulate a contextual response. The generation of multimodal content can increase this aspect – and the credibility of the content according to the MAIN model.

However, the content's credibility depends not only on the quality of the generated content. Although today's systems often generate convincing artifacts that are no longer identifiable as artificial even to humans, they do not always function flawlessly and convincingly. Still, this is less of a problem than we would assume in a classic communication situation (face-to-face) for several reasons:

First, as presented before, it is essential to consider the environment and the lack of external influences during communication in social media while at the same time including human scripts to deal with this type of communication. Content generated by automata is not presented entirely to a single user, but distributed to multiple users, so that no single user can see the whole picture of a campaign. Spelling mistakes in single messages could be perceived as simple slips and nonsensical posts, as human trolling, or misunderstanding from the receiver's side. Humans are programmed to understand messages sent by a communication partner not only by content but also by interpreting social cues like facial expressions or gestures. Consequently, communication with no or reduced social cues leads to misunderstandings since it is abstract and less intuitive. An exciting field of self-experience of this phenomenon has undoubtedly been text-based communication during the pandemic: the restricted environment of a chat platform may lead to frequent misunderstandings. Even using additional cues, such as emojis or stickers, is insufficient to solve this problem since they sometimes may even increase misunderstandings if a meaning of an emoji is ambiguous. An automaton must only act similar as humans would with all their errors and deviations to disguise its true identity from the message receiver.

Second, as mentioned in the perspective of content generation in games and social media, users believe that the things they are confronted with on social media are real. In contrast to gaming, where gamers are in a clearly virtual setting and pushed to the content and action to reach a suspension of disbelief, social media users belief in a real social setting and seemingly often do not want to scrutinize the origin of the information. Partly this may be the case since it is laborious to review every source in such a broad ecosystem like social media (similar to the situation of real social interaction scenarios). Additionally, tedious fact checking would pop the bubble on their social media platforms, where they can see personalized content which correspond to their world view. Especially in

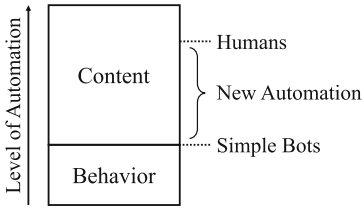

Fig. 1. Visualization of the gap between simple automatons and human communication that must be closed by New Automation techniques.

times of events like the pandemic, users rely on their social interactions via social media, making it very painful to disintegrate the reliability of these platforms for every single user. Thus, only strong cues that irritate users seem to be strong enough to make users question the validity of the information.

However, most of the limited cues transmitted via social media are enough to trigger humans' socialization scripts (e.g., as in the CASA model [49]) for ensuring believable communication. This offers two new levels of individual influence: (1) Cues can be reduced in a way that messages become more ambiguous (e.g., the use of fewer emojis will lead to uncertainty regarding the intent of the message and may leave a communication partner being insecure and occupied with deciphering the reduced message). Here, automation can easily be applied since it only needs to be sophisticated enough to provide enough cues such that users are not entirely sure whether their communication partner is an automaton or a human being. (2) Cues can be inserted to trigger specific social scripts for causing specific reactions. This is certainly more difficult and directly correlates to the sophistication of the applied content generation technology. Still, the auto-generated text has only been adjusted to the situation and should be good enough to fulfill human expectations, but not better so as to avoid triggering the Uncanny Valley.

As shown in Fig. 1 for the application of persuasive automation in online media, the challenge is not to accurately replicate interpersonal communication. It is to properly control cues and content in the setting of feature-poor communication in technical environments for creating sufficient uncertainty in the communication partner about the actual nature of an actor behind an account or avatar. New Automation has to close the gap between only simulating human behavior by producing content for completing the human appearance of an account or avatar in the restricted scenario of online media.

5 Future Challenges Implied by New Automation

The recent advancements to the tools and techniques capable of generating artificial content and driving the next generation of automated accounts pose opportunities, questions, and dire challenges. Among them are the challenges related to detecting AI-powered social bots, assessing the effects of New Automation,

measuring the quality of content and detecting low-quality one, designing and applying corrective interventions, and ethics.

5.1 Detection of Automation

AI-powered accounts could post multimedia content with human-like patterns by combining the capabilities of AI systems that generate realistic and credible behaviors, photos, videos, and texts. An account with these characteristics and whose behavior is decided by an AI to minimize its detectability while maximizing its impact inevitably poses much increased challenges than those faced by bot detectors up to now. This observation raises an important question as to whether *it will even be possible to distinguish such bots from human-operated accounts* in the future. In more than a decade of research on social bot detection, we witnessed countless efforts aimed at developing detectors capable of effectively spotting the majority of existing bots. This considerable effort led to the development of literally hundreds of different bot detectors [14]. Unfortunately, existing benchmark studies demonstrated the inherent difficulty of this task, which, as of now, still stands as largely unsolved [15,21,52,56]. The unsatisfactory results obtained against unresourceful bots cast a shadow on our capacity to detect future intelligent bots. To turn the tide in the fight against automated and other malicious accounts, some scholars proposed alternative approaches, such as those aimed at detecting coordinated behaviors rather than automated ones [50,77], or those that take into account the presence of adversaries by design [16]. Others, however, deemed the task too difficult and recommended policy, legal, and normative interventions to curb the many possible malicious applications of automation and AI in online media [10]. New Automation thus introduces a conundrum within this context: Detecting the next generation of social bots might prove simply too tricky or outright impossible, but leaving them be would make us vulnerable to their manipulations.

5.2 Measurement of Content Quality

An interesting and undoubtedly complex challenge in the context of New Automation is measuring content quality concerning a given context. Here, different measures were developed in the past, focusing on assessing the adequacy, fluency, diversity and factuality of the automatically generated texts [12]. Besides the complexity of finding suitable proxies for assessing these criteria in the multimodal domain [3], it is also a double-edged sword. On the one hand, assessing content under investigation needs to examine it for coherence with the broader context. On the other hand, using such a measure would be easy to identify incoherent content and poor combinations of multimodal constructs (e.g., image and text). At the same time, these measures would also be suitable to be used as optimization criteria for generating processes and thus for their improvement.

However, currently, there are no such combined measures available. Although some indicators for text quality exist, they do not measure what needs to be measured to judge artificially generated content in more than one dimension. While

the so-called BLEU Score [51] and its successors were initially being developed for evaluating machine translations (the closer translation to professional human translation, the better), the ROUGE score [42] was developed based on BLEU for text summarization. It compares the summary with the original text and implements different score versions (e.g., based on the longest common sub-strings or different numbers of n-grams as a basis). At least the BERT-Score [82] calculates (in contrast to BLEU and ROUGE) a semantic similarity score for each token in a candidate sentence with each token in a reference sentence. However, (a) all measures need a reference to compare with; (b) they cannot evaluate whether a text represents a specific opinion or a whether the text makes any sense in specific content, and (c) they only measure one specific aspect of the text's quality instead of providing an overall picture.

The only currently available option to check for the quality of the generated text is the evaluation by humans. As we have seen from the discussion of the New Automation paradigm in social media, this may not necessarily influence the applicability of the automation side but certainly the detection side. While humans may activate their social scripts to integrate artificial content into the current context, detection methods will fail to notice discrepancies objectively.

5.3 Effects of Automation

The effects derived from New Automation can be either positive or negative, depending on the use case, context and intentions. Positive effects may be derived from the increased communication efficiency. For example, if suitable methods have been designed that can detect the spread of fake news, content moderators may intervene early in the distribution process. Further, in particular situations like natural disasters, information can spread faster and be targeted more directly to the affected people. Additionally, although research may not be able to detect social bots anymore, it will maybe focus on mitigating the effects of their actions. Thus, the final goal – making social media an uninfluenced platform for the free exchange of opinions – may be achieved nevertheless.

However, the dark side of New Automation includes the scenario of information warfare [19]. If social bots and disinformation cannot be detected reliably, moderators or other concerned parties may use other methods like pre- or debunking to counteract these developments. This would increase the amount of content on social media, possibly one half in favor and the other half against a particular opinion. Overall, this polarized situation would decrease users' trust and reliance on social media. Especially in times of a pandemic, where many people are socially isolated, this may have severe psychological consequences. Lastly, if more and more content is posted online (and the creation of this content is not effortful anymore), communication itself may become arbitrary. Like industrialization decreased the efforts to create objects, making them more expendable, will the automation of word generation make conversations less valuable? Artificial content may eventually even dominate social interaction. If such data is used as input for training language generating models (as it is done currently on large corpora of text from the web), a self-enforcing cycle of stereotype language

generation may result. Whatever effects materialize, it seems to be certain that the nature and the intention of communication but maybe also New Automation itself will be affected by New Automation.

5.4 Moderation Interventions and New Platforms

A significant challenge in the face of New Automation is content moderation and moderation interventions – i.e., taking action directly and in a timely manner in ongoing events to stop abuses [28]. Simple regulation is not enough; automatic methods for detecting and contrasting automation, low-quality content and misbehavior must be implemented. At the level of the platform operators, this would mean permanent monitoring of data and content, which indeed harbors its dangers (for example, the censorship regulations of platform operators may damage their public image). Nevertheless, if one wants to take this path, the methodological gap in evaluating and classifying content exists, as described above. Then, in addition to simply detecting problematic content and behavior, effective content moderation also implies the deployment of adequate corrective actions (i.e., moderation interventions) [37,73]. The ultimate goal of moderation interventions is that of persuading users to drop harmful or otherwise problematic behavior (e.g., posting offensive or fake content). As such, applying moderation interventions automatically brings us back to the challenges of computer-mediated communication and of creating convincing AI-generated interventions (e.g., messages [9,70]). The design, (automated) deployment and evaluation of moderation interventions is still a relatively little explored area of research [17], and even more so in relation to New Automation.

At the same time, the human scripts and behaviors described above may provide a starting point in the long term for shifting attention away from these (instinctive) scripts and toward a critical approach to the content consumed. A first step could be to make it clear to users through the virtual environment design that social media are not a reflection of natural social interaction. Another step may be to warn users more often and openly about the difficulties of detecting, for example, social bots. An opposite trend will undoubtedly be the merging of virtual environments and social media in the next few years [20]. The so-called Metaverse could play an essential role in this. Users are undoubtedly aware that they are in a parallel, virtual world in this environment. It would be conceivable that in such an environment, the game world's rules dominate, creating a decoupling of virtual (and very global) reality and genuine social (often local) interaction. This makes it more challenging to transfer narratives and deception from the virtual world (including so-called extended reality) to the real world.

5.5 Ethical Implications

Finally, we want to briefly address several ethical issues that arise with research in the mentioned challenges but also with this paper itself. Any advancement in technology can be used for the prevention of malicious actions or applied in the context of malicious use (e.g., as part of the manipulation of disinformation

campaigns). This is true for measures of content quality, detection mechanisms and contrasting actions [17]. If misuse can be detected, countermeasures can be evaluated with these detectors. As such, automation and bot detection is in a continuous arms race with malicious actors that try to avoid detection [16].

However, this work not only contributes to a multifaceted perspective of possible near- and midterm developments in automated communication in online media but can also be understood as an invitation or idea provider for malicious actors to increase focus on human perception and new technologies as an effective entity in the context of New Automation. Nevertheless, we think it is more important to highlight the challenges and possible upcoming technology leaps implied by New Automation than to ignore the possibilities or even dismiss them as a conspiracy and hope for the best.

6 Conclusion

In this work, we have theoretically explored the topic of AI-driven New Automation of communication in social media under the use of modern generation technologies at the content level. To this end, three relevant perspectives have been incorporated: the research on automated (often very simple) communication in social media, the technological perspective on automated content generation, and the facet of automated content generation in games. Placement in existing models such as computer-mediated communication (CMC), computers are social actors (CASA), Uncanny Valley, and the Modality, Agency, Interactivity, and Navigability model (MAIN) allow us to predict that already current AI-based content generation technologies (such as GPT, DALL-E, or MAGMA) have sufficient capabilities to deceive human actors when communicating with automata (which hide behind abstract social media accounts). On the one hand, this deception is based on the very specific environment of social media - a very restricted environment in which important cues of human interaction are missing to make a confident statement about the counterpart. On the other hand, the very deliberate setting of cues can sow uncertainty about the nature of the counterpart, activating human interaction scripts and thus supporting a humanization of automata and with it also the acceptance of generated content.

The present work is theoretical in nature and is based on an analysis of existing current technologies which, at least according to the literature and the authors' state of knowledge, are not yet in widespread use. Therefore, there seems to be no need to speculate about coming technology leaps and their effects, as long as already the presented New Automation brings a large amount of challenges: Challenges in detecting advanced automation, measuring content quality, exploring the effects of New Automation, and the possibilities of corrective interventions by platforms.

References

1. Adelani, D.I., Mai, H., Fang, F., Nguyen, H.H., Yamagishi, J., Echizen, I.: Generating sentiment-preserving fake online reviews using neural language models and their human- and machine-based detection. In: Barolli, L., Amato, F., Moscato, F., Enokido, T., Takizawa, M. (eds.) AINA 2020. AISC, vol. 1151, pp. 1341–1354. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-44041-1_114
2. Alabdulkarim, A., Li, S., Peng, X.: Automatic story generation: challenges and attempts (2021). https://doi.org/10.48550/ARXIV.2102.12634. https://arxiv.org/abs/2102.12634
3. Alam, F., et al.: A survey on multimodal disinformation detection. In: The 29th International Conference on Computational Linguistics (COLING 2022). ACL (2022)
4. Alhayan, F., Pennington, D.R., Ruthven, I.: "She seems more human": understanding twitter users' credibility assessments of dementia-related information. In: Smits, M. (ed.) Information for a Better World: Shaping the Global Future. LNCS, vol. 13193, pp. 292–313. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-96960-8_20
5. Arsenyan, J., Mirowska, A.: Almost human? A comparative case study on the social media presence of virtual influencers. Int. J. Hum. Comput. Stud. **155**, 102694 (2021). https://doi.org/10.1016/j.ijhcs.2021.102694
6. Assenmacher, D., Clever, L., Frischlich, L., Quandt, T., Trautmann, H., Grimme, C.: Demystifying social bots: on the intelligence of automated social media actors. Soc. Media + Soc. **6**(3) (2020). https://doi.org/10.1177/2056305120939264
7. Assenmacher, D., et al.: Benchmarking crisis in social media analytics: a solution for the data-sharing problem. Soc. Sci. Comput. Rev. (2021). https://doi.org/10.1177/08944393211012268
8. Bessi, A., Ferrara, E.: Social bots distort the 2016 U.S. Presidential election online discussion. First Monday **21**(11) (2016). https://doi.org/10.5210/fm.v21i11.7090. https://firstmonday.org/ojs/index.php/fm/article/view/7090
9. Bilewicz, M., et al.: Artificial intelligence against hate: intervention reducing verbal aggression in the social network environment. Aggress. Behav. **47**(3), 260–266 (2021)
10. Boneh, D., Grotto, A.J., McDaniel, P., Papernot, N.: How relevant is the Turing test in the age of sophisbots? IEEE Secur. Priv. **17**(6), 64–71 (2019)
11. Brown, T., et al.: Language models are few-shot learners. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.) Advances in Neural Information Processing Systems, vol. 33, pp. 1877–1901. Curran Associates, Inc. (2020)
12. Chen, D., Dolan, W.: Collecting highly parallel data for paraphrase evaluation. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, Oregon, USA, pp. 190–200. Association for Computational Linguistics (2011). https://aclanthology.org/P11-1020
13. Ciechanowski, L., Przegalinska, A., Magnuski, M., Gloor, P.: In the shades of the uncanny valley: an experimental study of human-chatbot interaction. Futur. Gener. Comput. Syst. **92**, 539–548 (2019). https://doi.org/10.1016/j.future.2018.01.055
14. Cresci, S.: A decade of social bot detection. Commun. ACM **63**(10), 72–83 (2020)
15. Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., Tesconi, M.: The paradigm-shift of social spambots: evidence, theories, and tools for the arms race. In: The 26th International Conference on World Wide Web Companion (WWW 2017), pp. 963–972 (2017)

16. Cresci, S., Petrocchi, M., Spognardi, A., Tognazzi, S.: The coming age of adversarial social bot detection. First Monday **26**(7) (2021)
17. Cresci, S., Trujillo, A., Fagni, T.: Personalized interventions for online moderation. In: The 33rd ACM Conference on Hypertext and Social Media (HT 2022), pp. 248–251. ACM (2022)
18. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, Minnesota, pp. 4171–4186. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/N19-1423
19. Di Pietro, R., Caprolu, M., Raponi, S., Cresci, S.: New Dimensions of Information Warfare. Advances in Information Security, vol. 84. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-60618-3
20. Di Pietro, R., Cresci, S.: Metaverse: security and privacy issues. In: The 3rd IEEE International Conference on Trust, Privacy and Security in Intelligent Systems, and Applications (TPS 2021), pp. 281–288. IEEE (2021)
21. Echeverría, J., De Cristofaro, E., Kourtellis, N., Leontiadis, I., Stringhini, G., Zhou, S.: LOBO: evaluation of generalization deficiencies in Twitter bot classifiers. In: The 34th Annual Computer Security Applications Conference (ACSAC 2018), pp. 137–146 (2018)
22. Edwards, C., Beattie, A.J., Edwards, A., Spence, P.R.: Differences in perceptions of communication quality between a twitterbot and human agent for information seeking and learning. Comput. Hum. Behav. **65**, 666–671 (2016). https://doi.org/10.1016/j.chb.2016.07.003
23. Edwards, C., Edwards, A., Spence, P.R., Shelton, A.K.: Is that a bot running the social media feed? Testing the differences in perceptions of communication quality for a human agent and a bot agent on twitter. Comput. Hum. Behav. **33**, 372–376 (2014). https://doi.org/10.1016/j.chb.2013.08.013
24. Eichenberg, C., Black, S., Weinbach, S., Parcalabescu, L., Frank, A.: Magma-multimodal augmentation of generative models through adapter-based finetuning. arXiv preprint arXiv:2112.05253 (2021). https://doi.org/10.48550/arXiv.2112.05253
25. Fagni, T., Falchi, F., Gambini, M., Martella, A., Tesconi, M.: TweepFake: about detecting deepfake tweets. PLoS ONE **16**(5), e0251415 (2021)
26. Ferrara, E., Varol, O., Davis, C., Menczer, F., Flammini, A.: The rise of social bots. Commun. ACM **59**(7), 96–104 (2016)
27. Gallwitz, F., Kreil, M.: The rise and fall of 'social bot' research. SSRN 3814191 (2021). https://ssrn.com/abstract=3814191
28. Gillespie, T.: Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media. Yale University Press, New Haven (2018)
29. Grimme, C., Assenmacher, D., Adam, L.: Changing perspectives: is it sufficient to detect social bots? In: Meiselwitz, G. (ed.) SCSM 2018. LNCS, vol. 10913, pp. 445–461. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-91521-0_32
30. Grimme, C., Preuss, M., Adam, L., Trautmann, H.: Social bots: human-like by means of human control? Big Data **5**(4), 279–293 (2017). https://doi.org/10.1089/big.2017.0044
31. Groover, M.: Fundamentals of Modern Manufacturing: Materials, Processes, and Systems. Wiley, Hoboken (2010)

32. He, B., Ahamad, M., Kumar, S.: PETGEN: personalized text generation attack on deep sequence embedding-based classification models. In: The 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (KDD 2021), pp. 575–584 (2021)
33. Ho, A., Hancock, J., Miner, A.S.: Psychological, relational, and emotional effects of self-disclosure after conversations with a chatbot. J. Commun. **68**(4), 712–733 (2018). https://doi.org/10.1093/joc/jqy026
34. Im, J., Tandon, S., Chandrasekharan, E., Denby, T., Gilbert, E.: Synthesized social signals: computationally-derived social signals from account histories. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA. ACM (2020). https://doi.org/10.1145/3313831.3376383
35. Ippolito, D., Duckworth, D., Callison-Burch, C., Eck, D.: Automatic detection of generated text is easiest when humans are fooled. In: Proceedings of the 58th Annual Meeting of the ACL, pp. 1808–1822. ACL (2020). https://doi.org/10.18653/v1/2020.acl-main.164
36. Jeon, Y.A.: Reading social media marketing messages as simulated self within a metaverse: an analysis of gaze and social media engagement behaviors within a metaverse platform. In: 2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW), pp. 301–303 (2022). https://doi.org/10.1109/VRW55335.2022.00068
37. Jhaver, S., Boylston, C., Yang, D., Bruckman, A.: Evaluating the effectiveness of deplatforming as a moderation strategy on Twitter. In: The 24th ACM Conference On Computer-Supported Cooperative Work And Social Computing (CSCW 2021). ACM (2021)
38. Johnson, M.R.: Collecting highly parallel data for paraphrase evaluation. In: Proceedings of the 6th Workshop on Procedural Content Generation. Dundee, Scotland, UK (2016). https://www.pcgworkshop.com
39. Knox, J.: The metaverse, or the serious business of tech frontiers. Postdigit. Sci. Educ. **4**(2), 207–215 (2022). https://doi.org/10.1007/s42438-022-00300-9
40. Lee, L.H., et al.: All One Needs to Know about Metaverse: A Complete Survey on Technological Singularity, Virtual Ecosystem, and Research Agenda (2021). arXiv Preprint. https://doi.org/10.48550/ARXIV.2110.05352
41. Liapis, A., Yannakakis, G.N., Nelson, M.J., Preuss, M., Bidarra, R.: Orchestrating game generation. IEEE Trans. Games **11**(1), 48–68 (2019). https://doi.org/10.1109/TG.2018.2870876
42. Lin, C.Y.: ROUGE: a package for automatic evaluation of summaries. In: Text Summarization Branches Out, Barcelona, Spain, pp. 74–81. ACL (2004). https://aclanthology.org/W04-1013
43. Livingstone, D.: Turing's test and believable AI in games. Comput. Entertain. **4**(1), 6 (2006). https://doi.org/10.1145/1111293.1111303
44. Mendoza, M., Tesconi, M., Cresci, S.: Bots in social and interaction networks: detection and impact estimation. ACM Trans. Inf. Syst. **39**(1), 1–32 (2020)
45. Meta Platforms: The Facebook Company Is Now Meta (2021). https://about.fb.com/news/2021/10/facebook-company-is-now-meta/. Accessed 22 May 2022
46. Mnih, V., et al.: Human-level control through deep reinforcement learning. Nature **518**(7540), 529–533 (2015). https://doi.org/10.1038/nature14236
47. Mori, M.: The uncanny valley. Energy **7**(4), 33–35 (1970). https://doi.org/10.1109/MRA.2012.2192811
48. Mozgovoy, M., Preuss, M., Bidarra, R.: Guest editorial special issue on team AI in games. IEEE Trans. Games **13**(4), 327–329 (2021). https://doi.org/10.1109/TG.2021.3127967

49. Nass, C., Steuer, J., Tauber, E.R.: Computers are social actors. In: Conference Companion on Human Factors in Computing Systems, Boston, MA, USA, CHI 1994, p. 204. Association for Computing Machinery (1994). https://doi.org/10.1145/259963.260288
50. Nizzoli, L., Tardelli, S., Avvenuti, M., Cresci, S., Tesconi, M.: Coordinated behavior on social media in 2019 UK general election. In: The 15th International AAAI Conference on Web and Social Media (ICWSM 2021), pp. 443–454. AAAI (2021)
51. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL 2002, Philadelphia, Pennsylvania, p. 311. Association for Computational Linguistics (2001). https://doi.org/10.3115/1073083.1073135
52. Pohl, J.S., Assenmacher, D., Seiler, M.V., Trautmann, H., Grimme, C.: Artificial social media campaign creation for benchmarking and challenging detection approaches. In: Proceedings of the 16th International Conference on Web and Social Media. NEATCLasS, Association for the Advancement of Artificial Intelligence (AAI), Hybrid: Atlanta, Georgia, US and Online (2022)
53. Rabkin, M.: Connect 2021 Recap: Horizon Home, the Future of Work, Presence Platform, and More (2021). https://www.oculus.com/blog/connect-2021-recap-horizon-home-the-future-of-work-presence-platform-and-more/. Accessed 22 May 2022
54. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding by generative pre-training. Technical report, OpenAI (2018)
55. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 (2022). https://doi.org/10.48550/ARXIV.2204.06125
56. Rauchfleisch, A., Kaiser, J.: The false positive problem of automatic bot detection in social science research. PLoS ONE **15**(10), e0241045 (2020)
57. Rauschnabel, P.A., Felix, R., Hinsch, C., Shahab, H., Alt, F.: What is XR? Towards a framework for augmented and virtual reality. Comput. Hum. Behav. **133**, 107289 (2022). https://doi.org/10.1016/j.chb.2022.107289
58. Reed, S., et al.: A generalist agent (2022). https://doi.org/10.48550/ARXIV.2205.06175. https://arxiv.org/abs/2205.06175
59. Reeves, B., Nass, C.: The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Pla. Bibliovault OAI Repository, the University of Chicago Press (1996)
60. Riedl, M.O.: Computational narrative intelligence: a human-centered goal for artificial intelligence. arXiv preprint arXiv:1602.06484 (2016)
61. Risi, S., Preuss, M.: From chess and atari to StarCraft and beyond: how game AI is driving the world of AI. KI - Künstliche Intelligenz **34**(1), 7–17 (2020). https://doi.org/10.1007/s13218-020-00647-w
62. Saygin, A.P., Chaminade, T., Ishiguro, H., Driver, J., Frith, C.: The thing that should not be: predictive coding and the uncanny valley in perceiving human and humanoid robot actions. Soc. Cogn. Affect. Neurosci. **7**(4), 413–422 (2012). https://doi.org/10.1093/scan/nsr025
63. Shin, M., Song, S.W., Chock, T.M.: Uncanny valley effects on friendship decisions in virtual social networking service. Cyberpsychol. Behav. Soc. Netw. **22**(11), 700–705 (2019). https://doi.org/10.1089/cyber.2019.0122
64. Silver, D., et al.: Mastering the game of go with deep neural networks and tree search. Nature **529**(7587), 484–489 (2016). https://doi.org/10.1038/nature16961

65. Skjuve, M., Haugstveit, I., Følstad, A., Brandtzaeg, P.: Help! Is my chatbot falling into the uncanny valley? An empirical study of user experience in human-chatbot interaction. Hum. Technol. **15**, 30–54 (2019). https://doi.org/10.17011/ht/urn.201902201607
66. Spence, P.R., Edwards, A., Edwards, C., Jin, X.: 'The bot predicted rain, grab an umbrella': few perceived differences in communication quality of a weather twitterbot versus professional and amateur meteorologists. Behav. Inf. Technol. **38**(1), 101–109 (2019). https://doi.org/10.1080/0144929X.2018.1514425
67. Stephenson, N.: Snow Crash. Metropolis Media (1992)
68. Sundar, S.S.: The MAIN Model: A Heuristic Approach to Understanding Technology Effects on Credibility. Digital Media, p. 29 (2008)
69. Tardelli, S., Avvenuti, M., Tesconi, M., Cresci, S.: Characterizing social bots spreading financial disinformation. In: Meiselwitz, G. (ed.) HCII 2020. LNCS, vol. 12194, pp. 376–392. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-49570-1_26
70. Tekiroglu, S., Bonaldi, H., Fanton, M., Guerini, M.: Using pre-trained language models for producing counter narratives against hate speech: a comparative study. In: Findings of the Association for Computational Linguistics (ACL 2022), pp. 3099–3114. ACL (2022)
71. Togelius, J.: We tried learning AI from games. How about learning from players? (2022). https://modl.ai/learning-ai-from-players. modl.ai blog
72. Togelius, J., et al.: Procedural content generation: goals, challenges and actionable steps. In: Lucas, S.M., Mateas, M., Preuss, M., Spronck, P., Togelius, J. (eds.) Artificial and Computational Intelligence in Games, Dagstuhl Follow-Ups, vol. 6, pp. 61–75. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany (2013). https://doi.org/10.4230/DFU.Vol6.12191.61. http://drops.dagstuhl.de/opus/volltexte/2013/4336
73. Trujillo, A., Cresci, S.: Make reddit great again: assessing community effects of moderation interventions on r/The_Donald. In: The 25th ACM Conference On Computer-Supported Cooperative Work And Social Computing (CSCW 2022). ACM (2022)
74. Tsimpoukelli, M., Menick, J., Cabi, S., Eslami, S.M.A., Vinyals, O., Hill, F.: Multimodal few-shot learning with frozen language models. In: Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems (2021)
75. Vaswani, A., et al.: Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS 2017, pp. 6000–6010. Curran Associates Inc., Red Hook (2017)
76. Walther, J.B.: Computer-mediated communication: impersonal, interpersonal, and hyperpersonal interaction. Commun. Res. **23**(1), 3–43 (1996). https://doi.org/10.1177/009365096023001001
77. Weber, D., Neumann, F.: Amplifying influence through coordinated behaviour in social networks. Soc. Netw. Anal. Min. **11**(1), 1–42 (2021). https://doi.org/10.1007/s13278-021-00815-2
78. Xu, K., Liao, T.: Explicating cues: a typology for understanding emerging media technologies. J. Comput.-Mediat. Commun. **25**(1), 32–43 (2020). https://doi.org/10.1093/jcmc/zmz023
79. Yang, C., Harkreader, R., Gu, G.: Empirical evaluation and new design for fighting evolving Twitter spammers. IEEE Trans. Inf. Forensics Secur. **8**(8), 1280–1293 (2013)

80. Yang, K.C., Varol, O., Hui, P.M., Menczer, F.: Scalable and generalizable social bot detection through data selection. In: Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, vol. 34 (2020). https://doi.org/10.1609/aaai.v34i01.5460
81. Yannakakis, G.N., Togelius, J.: Experience-driven procedural content generation. IEEE Trans. Affect. Comput. **2**(3), 147–161 (2011). https://doi.org/10.1109/T-AFFC.2011.6
82. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: BERTScore: evaluating text generation with BERT. arXiv preprint arXiv:1904.09675 (2019)