



HAL
open science

Pay Attention: a Call to Regulate the Attention Market and Prevent Algorithmic Emotional Governance

Franck Michel, Fabien Gandon

► To cite this version:

Franck Michel, Fabien Gandon. Pay Attention: a Call to Regulate the Attention Market and Prevent Algorithmic Emotional Governance. AIES 2024 - 7th AAAI/ACM Conference on AI, Ethics, and Society, AAAI/ACM, Oct 2024, San Jose (CA), United States. hal-04758276

HAL Id: hal-04758276

<https://hal.science/hal-04758276v1>

Submitted on 29 Oct 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

Pay Attention: a Call to Regulate the Attention Market and Prevent Algorithmic Emotional Governance

Franck Michel¹, Fabien Gandon²

¹University Côte d’Azur, CNRS, Inria

²University Côte d’Azur, Inria, CNRS

franck.michel@inria.fr, fabien.gandon@inria.fr

Abstract

Over the last 70 years, we, humans, have created an economic market where attention is being captured and turned into money thanks to advertising. During the last two decades, leveraging research in psychology, sociology, neuroscience and other domains, Web platforms have brought the process of capturing attention to an unprecedented scale. With the initial commonplace goal of making targeted advertising more effective, the generalization of attention-capturing techniques and their use of cognitive biases and emotions have multiple detrimental side effects such as polarizing opinions, spreading false information and threatening public health, economies and democracies. This is clearly a case where the Web is not used for the common good and where, in fact, all its users become a vulnerable population. This paper brings together contributions from a wide range of disciplines to analyze current practices and consequences thereof. Through a set of propositions and principles that could be used to drive further works, it calls for actions against these practices competing to capture our attention on the Web, as it would be unsustainable for a civilization to allow attention to be wasted with impunity on a world-wide scale.

1 An Unsustainable Attention Market

Since the advent of mass market in the 50’s, media and advertisement providers have relentlessly tried to figure out effective methods to capture our attention and turn it into revenue. During the last two decades, supported by advances in artificial intelligence (AI), major online social media and Web platforms have brought this process of capturing attention to an unprecedented scale. Based almost exclusively on advertising revenues, their business model consists in providing free services that, in return, collect behavioral traces. This data is then used to maximize the impact of advertisements on users by ① ensuring their mental availability at the time of being shown the advertisement, and ② ensuring that the message meets their interests, beliefs and moods (i.e. targeted advertising). Based on research in psychology, sociology and neuroscience, several actors including online social media, games and Web platforms have engineered techniques capable of very effectively plundering our “available brain time” (Lewis 2017; Harris 2019). We can distinguish two broad categories of such techniques.

Firstly, some techniques are explicitly designed to **leverage cognitive biases** as a means to capture attention. For in-

stance, the *likes* collected after posting content activate the brain’s dopaminergic pathways (involved in the reward system) and tap into our need for social approval, giving “bright dings of pseudo-pleasure” (Lewis 2017); notifications of smartphone applications feed our appetite for novelty and surprise such that it is difficult to resist the urge to check them; the pull-to-refresh mechanism (Lewis 2017), alike slot machines, exploits the variable reward pattern whereby each time we pull down the screen we may get an update or nothing at all; infinite scrolling (of news, posts or videos...) traps us because of our fear of missing out important information (FOMO) to the point that we can hardly break the flow; automatic video chaining replaces a deliberate action to continue watching with a required action to stop watching, and entails a frustrating feeling of incompleteness when stopped; etc. Similarly, some techniques harness dark patterns¹ (Gray et al. 2023) to manipulate users into taking actions or decisions they wouldn’t take otherwise. This is typically the case when one accepts all notifications of an application without really noticing it, while deactivating notifications would require an additional, less intuitive, series of actions.

Secondly, recent advances in machine learning allow the training of **content recommendation algorithms** on massive online behavioral data, which Zuboff refers to as “behavioral surplus” (Zuboff 2019). These algorithms learn to recommend content that not only captures attention but also increases user engagement². They discover the content’s key features that help predict whether such content will effectively attract users’ attention, and typically end up selecting content related to conflictuality, fear or sexuality (Bronner 2021). They also learn to exploit humans’ negativity bias (Soroka, Fournier, and Nir 2019; Siegrist and

¹The legal definition in California is “A user interface is a dark pattern if the interface has the effect of substantially subverting or impairing user autonomy, decision-making, or choice. A business’s intent in designing the interface is not determinative in whether the user interface is a dark pattern, but a factor to be considered.” CPRA § 7004 (c)

²There are multiple definitions of user engagement. In the context of social media, this typically refers to the fact that a user would interact with a content: e.g. like, comment or repost it. Engagement is usually public in that it leaves public traces on the platform, unlike sheer content consumption that remains private (Robertson et al. 2023).

Cvetkovich 2001) and, as a consequence, content conveying high-arousal negative emotions (such as anger, resentment, indignation and disgust) are more likely to be read and eventually shared online than those conveying other emotions (Robertson et al. 2023; Kohout, Kruikemeier, and Bakker 2023). Concerningly enough, false information (a broad term including misinformation and other forms of disinformation) typically relies on such negative emotions as a trick to foster sharing. Finally, recommendation systems may do all this without it being explicit, neither in the features they select nor in the succinct feedback that some of them happen to provide users with³.

Since the amount of attention available is both limited and precious, it would be unsustainable for a civilization to waste it massively and with impunity for questionable or futile purposes (Bronner 2021). Today, we might precisely be at that moment: while mental time has become a new oil, we have created an attention economy and subsequent attention markets (Hefti and Heinke 2015; Hendricks and Vestergaard 2019) that, although sustainable from an economic point of view, may be unsustainable from a civilization point of view. From these first references, let us define what the term “attention market” refers to in this article.

Definition 1.1 (Attention Market) *Economic environment where businesses compete to capture and retain the resource represented by people’s focused mental engagement that we call attention.*

The attention market treats attention as a tradable commodity and involves multiple actors: from producers (the end users whose attention is the resource), to content creators whose work is used to capture the attention, brokers who trade and monetize the attention, and consumers (the large Web platforms) who use it for their own purposes such as exposing users to advertisements.



In this article, we propose a discussion aimed at spurring introspection and debate, primarily within the computer science community which designs these technologies, but also more generally within all the communities concerned about societal issues of AI technologies. Our goal is not to provide an additional piece of evidence, for there are many as we will see, but to insist on the interdisciplinary recognition of the situation, of its urgency and of the possible means of action. In line with the Web Science Manifesto (Berendt et al. 2021) calling for interdisciplinary approaches to prepare the future of the Web, we bring together and synthesize the conclusions of more than 70 papers and books from a wide range of disciplines to analyze the practices and drifts of these systems designed to capture attention on a worldwide scale. We make the point that, with the initial commonplace goal of making targeted advertising more effective, the generalization of attention-capturing techniques and their use of negative emotions tends to foster radicalization and polariza-

³For instance, a recommendation system may tell us “you liked this movie, you may also like this one”. But we don’t know what features were selected to recommend this one: Do they have an actor in common? Did my contacts like both of them? etc.

tion, amplify the dissemination of false information, spur the emergence of populism, and eventually put a threat on democracies and human societies in general.

If we look specifically at the computer science community, it appears that, so far, the public and private research has invested large efforts in dealing with some aspects of the problem like radicalization, violent speech and false information. Most of these works rely on *post hoc* measures (measures taken after the problems have appeared) such as content detection, deletion or downgrading (decreasing the content ranking so that it is less likely to be recommended). Nevertheless, we argue that additional measures must be considered to actively prevent the issues that stem from attention capturing rather than only mitigating their impact once they have occurred. Presumably, such measures would be political and economical as well as technical, meaning that this socio-technical problematic situation calls for socio-technical solutions. Hence this call for an interdisciplinary approach meant to regulate the attention market.

In the rest of this paper we will first review the general principles of recommendation systems and the consequences of the recommendation loop that they implement (section 2). Then, we will explain how having recommendation systems harness emotions can lead to detrimental situations including what we shall name an algorithmic emotional governance (section 3). We will touch upon the threat to creative jobs (section 4) and then review some known *post hoc* measures (section 5), before discussing preventive measures to reclaim our attention (section 6).

2 Users in the Loop... of Recommendation Systems

Content recommendation algorithms are a key component of a wide range of applications, including social media, search engines and major Web platforms in general. Through many applications they have changed our lives, helping us to be more efficient, assisting us in daily tasks, or improving our education and information. In a number of other applications however, the reality is not so bright. In the case of social media for instance, they are presented to us as if designed to provide us with content that matches our needs and desires, while what they really seek is to maximize the attention we pay to their hosting platform and advertisements thereof.

Through the training process, recommendation algorithms automatically learn to extract from massive behavioral traces the content’s features that most effectively capture our attention and maximize our continuous engagement with the platform. Typically, they can learn that some categories of topics, such as conflictuality, fear or sexuality, irresistibly attract our attention (Bronner 2021), and thus lean toward recommending these particular categories. They can also learn to select content tailored for a certain user, by taking into account both the content’s features (topics, source, emotions conveyed...) and its adequacy with the user’s profile (interests, inclinations, past behavior...). This adequacy likely involves many other features that are not transparent since the platforms rarely inform users about how and for

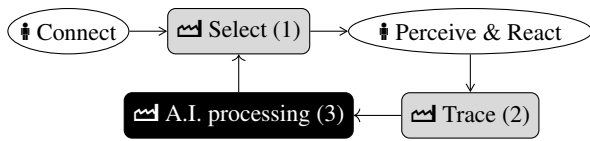


Figure 1: The self-reinforcing recommendation loop of platforms: the ellipses are activities on the user side, the boxes are activities on the platform side. *Select* and *Trace* are grey boxes because only partially observable. The A.I. processing is, more than often, a black box for the end-user.

which purpose their personalized feed was composed. This is underlined in a study by DeVito (DeVito 2017) who analyzed Facebook’s patents, press releases, and Securities and Exchange Commission filings, to identify “the set of algorithmic values that drive the News Feed”. Some of the features he identified are objective, i.e. they can be observed or measured: friend relationships, explicitly expressed user interests, prior engagement, post age and page relationships. By contrast, other features are up to interpretation and thus raise multiple questions: implicitly expressed user preferences (what are the signals of such implicit expression?), platform priorities (what are they and who decides them?), content quality (what are the quality criterion?).

Finally, it may seem that recommendation algorithms learn to leverage psychological traits and cognitive biases. Yet, it is important to stress that the algorithm does not discover such things as a psychological trait or a cognitive bias itself. Rather, it discovers the features that enable it to exploit what psychologists would describe as a trait or bias. Such criteria are not explicitly formulated, they may not even be explicable nor verifiable. They remain implicit in the models unless a study be carried out a posteriori, that would surface the biases that emerge from the recommendations. This is yet another example where AI techniques, without explanations nor feedback, raise ethical concerns.

Another specificity of recommendation algorithms is that they tend to implement a self-reinforcing loop that we define as follows:

Definition 2.1 (Self-Reinforcing Recommendation Loop)

The continuous cycle of recommendation systems providing personalized suggestions to a user based on data collected from their preferences and behaviors and integrating these to further recommendations.

A classical self-reinforcing recommendation loop is illustrated in figure 1: ① The algorithms recommend content to the user. ② The behavior of the user is captured, possibly partially due to the focus of the platform and the limited choices that the interface offers, and possibly biased due to the fact that these choices may be oriented, again by the interest of the platform and the chosen interface. ③ The algorithms integrate these reactions in future recommendations. As a result, the reactions of the user will reinforce the recommendation and propagation of the content.

Of course there are externalities to that loop, that can increase its impact. Smartphones, for example, provide additional means to profile users by tracking their every moves,

making recommendation even more efficient and targeted to the point that it competes and sometimes takes over more traditional ways of advertising (Times 2017). Another (detrimental) externality of this loop is that it opens the door to spoofing techniques and other malevolent actors intentionally biasing usage traces to “hijack” recommendation systems. Indeed, as soon as a process is known and documented, it runs the risk of being diverted from its original purpose and manipulated beyond its original objectives. For instance, fake reviews and reactions alter recommendations; black hat techniques of SEO (Search Engine Optimization) such as hidden texts, link farms, cloaking⁴ or text spinning are disapproved by search engines as they impact the recommendations they make by unduly increasing the ranking of targeted pages or avoiding their downgrading.



As a result, the fact that a few recommendation systems influence a significant fraction of the human population may have a number of detrimental side effects on their users and our societies at large. A first side effect is that recommendation algorithms tend to lock users in an informational space in accordance with their tastes and beliefs, a “filter bubble” (Pariser 2011) that confines them to a “cognitive comfort zone” and activates their confirmation bias as they are faced with information which seems to go towards the same directions or conclusions (Sasahara et al. 2021; Kitchens, Johnson, and Gray 2020). Eventually, users are no longer confronted with contradiction, debate nor disturbing facts or ideas, and this algorithmic amplification tends to be a powerful driver of the radicalization and polarization of opinions, leading to extremist ideas in some cases (Whittaker et al. 2021).

Furthermore, at a time where we need to change our behaviors (e.g. over-consumption of goods and energy) and redirect our attention to important matters (e.g. climate change), we should question whether recommendation algorithms make the right recommendations, and for whom. Considering the billions of users caught in recommendation loops everyday⁵, it is important to continuously monitor how and for what purpose these systems capture our attention. Because when our attention is spent on a content chosen by these platforms, it is lost for anything else.

3 Algorithmic Emotional Governance

Considering the platform’s recommendation loop introduced in section 2, we now want to stress that, directly or indirectly, emotions are a key feature of the selected recommendations. In fact, the whole attention market could be

⁴Cloaking denotes a technique in which the content presented to a search engine crawler is different from that presented to an actual user. It aims at deceiving search engines so they display the page that they would otherwise downgrade or dismiss. Adapted from <https://en.wikipedia.org/wiki/Cloaking>.

⁵In 2018, Google revealed that 70% of the time spent watching videos on Youtube is about videos recommended by Youtube’s algorithms. <https://qz.com/1178125/youtubes-recommendations-drive-70-of-what-we-watch>

seen as driven by a complex equation involving, at least, emotions, cognitive biases and content recommendation algorithms. This could lead to what Patino calls an “emocracy, a regime that makes our emotions become performative and invade public space” (Patino 2024). Here we will rather speak of an *algorithmic emotional governance* merging two concepts: emotional governance (Richards 2007) which is the informed management of the emotional dynamics of the governed population, and algorithmic governance (Rouvroy and Berns 2013) which is a governance of our societies based on the algorithmic processing of massive data.

Definition 3.1 (Algorithmic Emotional Governance) *The governance of societies based on algorithms processing massive data to harness the emotional dynamics of the governed population.*

Emotions are a powerful attractor of our attention, especially emotions with a high negative valence (Soroka, Fournier, and Nir 2019). As a result, information that arouses anger, fear, indignation, resentment, frustration or disgust is among those that most effectively catch our attention (Robertson et al. 2023; Kohout, Kruikemeier, and Bakker 2023). An explanation is that witnessing others’ negative emotions activates our comparison bias and subjects us to some sort of injunction to take sides, to show our emotional response, and hence publicly demonstrate our “irreproachable morality” (Crockett 2017; Bronner 2021). Note that catching attention and increasing user engagement are different things, and although high-arousal negative emotions catch attention more efficiently than other emotions, it remains unclear whether they induce a higher user engagement on social media. In some cases a higher sharing rate of information conveying positive emotions was observed (Keib et al. 2018; Kramer 2012). Nevertheless, in several contributions, researchers showed the overwhelming impact of emotions in argumentation and debates and the means to detect them (Benlamine et al. 2015; Villata et al. 2018; Basile et al. 2016), and it has also been shown that anger spreads faster on social media than any other emotion (Fan et al. 2014). Note that this attraction for negative content can be observed in completely different domains, e.g. in literature where the anti-utopian and dystopian fiction genres became more prominent within the utopian genre (Mascarell 2020).

Combined together, the construction of filter bubbles by recommendation systems and the ability of these systems to learn the content’s features that trigger a particular emotional response in a particular individual, can lead to some form of polarisation and end up trapping users in radicalization pathways. Consider the supporter of a sports club: it is because the system chooses the right topic (e.g. the right sport), the right content (e.g. an article about an opponent club) and the right tone and emotion (e.g. mocking criticism) that an emotion is provoked, followed by a registered reaction (like, comment, repost) and, over time, a potential polarisation is developed such as hatred for the opponent’s supporters.

Recommendation after recommendation, the filter bubble becomes an opinion bubble where users are isolated

from discrepant opinions, and eventually an emotion bubble where they are maintained in certain emotional states. In the end, the complex interaction of negative emotions, cognitive biases (e.g. negativity bias and impulsive tendency to show indignation) and recommendation algorithms leads to an emotional escalation. Often, this escalation is further worsened by the affordances offered by the platforms, that tend to make exchanges ever briefer and more simplistic: How to express a nuanced reflection in a 280-character tweet? How to underline a doubt when the only available choices are essentially limited to / (and sometimes)? How to agree with one part of a post and disagree with another one when this post is treated as a monolithic block by the interface that only offers the options //? Some examples of widgets from well-known platforms are illustrated in figure 2. This extreme discretization of choices adds to the mechanisms at work and reinforces the polarization of opinions and communities. Some dark patterns are even intentionally employed to make some actions easy and some more difficult: for instance, in Facebook the button to like a post is always visible whereas the option to report a post is at the bottom of a sub-menu, a pattern falling in the category known as “longer than necessary” (Board 2022).

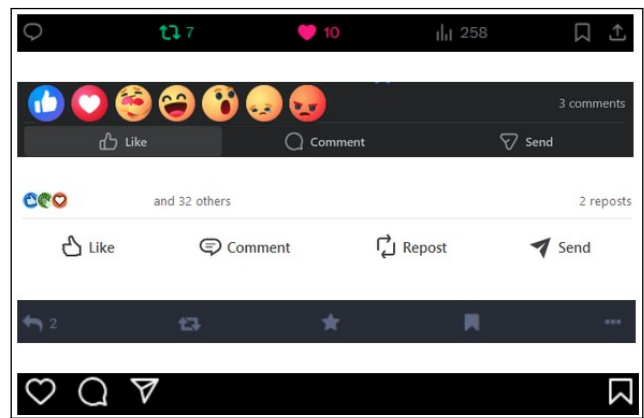


Figure 2: Five interfaces of well-known social media to react to a post.

Eventually, nuance, doubt or agnosticism are mechanically made invisible because the low emotional response that they induce simply downgrades their ranking. It is imperative to have an opinion, preferably definite and cleaving. Amplified by digital disinhibition⁶ (Suler 2004), this emotional escalation can lead to outpourings of violence and hatred whose outcome is sometimes tragic as attested by the suicide attempts of teenagers being cyberbullied (Schonfeld et al. 2023). Moreover, the full consequences of triggering or regulating emotions on our cognitive functions in general and on memory in particular remain to be studied extensively (Richards and Gross 2000).



We just described the combined effect of emotions, cognitive biases and recommendation algorithms, which is at

⁶the feeling of impunity induced by the feeling of anonymity

work whatever the type of content a platform serves. But things get even worse when it comes specifically to false information. False information are frequently meant to arouse strong negative emotions (Zaeem, Li, and Barber 2020), and the combination with cognitive biases and recommendation algorithms provides them with a particularly fertile ground and a formidable cognitive efficiency (Acerbi 2019; Martel, Pennycook, and Rand 2020). Some studies reported that negatively biased fake news enhance users' willingness to share them (Corbu et al. 2021), and reveal a positive correlation between the virality of fake news and the anger they carry (Chuai and Zhao 2022). Another study contended that falsehood spreads "significantly farther, faster, deeper, and more broadly than the truth" on social media (Vosoughi, Roy, and Aral 2018), which underlines that recommendation of content arousing negative emotions does not only induce local individual reaction: it creates a chain reaction leveraging the network effect of social media to spread that "content-emotion" couple through the acquaintance links. Other studies reported that recommendation algorithms mechanically tend to favor false information conveying divisive ideas, shocking events and negative emotions (Fernández, Bellogín, and Cantador 2021; Fortunato and Pecoraro 2022). This type of content entails a felt injunction to take sides and compulsively spread shocking information rather than appeal to critical thinking, question its veracity and verify its source. And since this information is often relayed by acquaintances, the social proof bias (Cialdini 2008) entices users to deem it credible and trustworthy.

Concerningly, the contents we are exposed to leave a trace in our implicit memory: although we cannot recall seeing it, it may impact our choices for several months (Courbet et al. 2014). Even more concerning is the fact that, due to the negativity bias, negative information leaves a longer memory trace than positive information. Therefore, even when a false information is denied or rectified, there remains a negative feeling that stems from the strong emotional response it triggered in the first place. Repeated again and again, associated with representations of the world that summon conspiracy theories, reinforced under the pressure of filter and emotion bubbles, and propelled by the network effect, such information tends to gradually and insidiously undermine our trust in experts (scholars, journalists, etc.), entails risks for public health (Waszak, Kasprzycka-Waszak, and Kubanek 2018; Rocha et al. 2023), and spurs the emergence of extreme ideas and populism that eventually undermine democracies (Whittaker et al. 2021; Abreu and França 2021; Intelligence 2022; Fortunato and Pecoraro 2022), among other pitfalls.

Finally, let us stress that if "previous studies have shown how personality, values, emotions and vulnerability of users affect their likelihood to propagate misinformation" (Fernández, Bellogín, and Cantador 2021), in this section we only considered an average user without any particular health condition. But we should envisage more complex situations when it comes to users with disabilities or mental disorders e.g. depression, anxiety, compulsive shopping disorder, paranoia, FOMO, FOBO⁷... Let us just mention one

⁷Fear of Better Options: the inability to choose when faced with

specific condition: the attention deficit (AD) disorder. There is evidence that AD symptoms could be worsened by the use of digital media and their attention-grabbing applications, and more importantly that these applications could provoke AD among people without previous record of such a disorder (Ra et al. 2018). To the very least, more research is needed in this respect.

4 Attention, Attention, all Thinkers

We firstly intended this section for all the scientists reading this paper, concurring with the article of David R. Smith: "Attention, attention: your most valuable scientific assets are under attack" (Smith 2018). In this article, Smith calls for attention to what media platforms are doing to research and the academic domain. Indeed, even the most informed scientists and engineers are not immune to these problems (Lewis 2017) such that digital contraptions (as Smith calls them) are contributing to *academic attention deficit disorder* (Smith 2018). In fact, concentration but also boredom, mind-wandering and daydreaming times are vital to creative thinking. Many of us experienced the sudden burst of an idea in the middle of a relaxing moment. Attention-capturing systems steal these moments from all of us and hamper the creativity process of wondering minds (Zomorodi 2017).

Of course these remarks can be generalized to many other activities requiring concentration, creativity and imagination, and one could wonder what digital contraptions are doing to politics, healthcare or education, for instance. To mention just one example, countless information media now report the cases of Youtubers experiencing a burnout (Parkin 2018), or musicians complaining that they spend more time making Tiktok videos to promote their music than actually creating music (Shah 2022; Whateley 2023). This reveals that, to hook and keep the attention of content consumers, platforms also exercise some sort of visibility tyranny over content creators.

In other words: attention, attention, thinkers, we need to redesign the systems for our own needs, rather than the other way around, especially in creative jobs since the true currency of these jobs are ideas (Smith 2018).

5 Known Post Hoc Counter-Measures

Among the various issues raised in the previous sections, the questions of false information, radicalization, hateful speech and bullying are among the most concerning, and therefore have been extensively addressed by the research community (Sharma et al. 2019). In (Fernández, Bellogín, and Cantador 2021) authors identified three different points where recommendation systems can be adapted to tackle these issues: ① pre recommendation, ② within the recommendation model, and ③ post recommendation. Most of the current counter-measures to deal with false information lie in this third category. Below we touch upon some of them.

Firstly, to dyke the spread of false information as well as inappropriate content such as bullying, hateful or violent speech, social media and content hosting platforms have

a multitude of options.

obligations that vary depending on the legislation and its jurisdiction (Funke and Flamini 2023). Measures range from content deletion and suspension of users spreading inappropriate content, to re-ranking of recommended items before presenting them to the user (Fernández, Bellogín, and Cantador 2021), flagging to indicate potentially deceptive content, etc. Yet, despite these various approaches, progress is still necessary. For instance, subtle violent content may be hard to detect as soon as it does not contain explicit hateful or violent terms, or when it uses sarcasm (Ocampo et al. 2023). Conversely, content may be erroneously assessed as abusive or illicit although it is in fact using irony to convey perfectly acceptable ideas. An in-depth analysis of implicit and subtlety in linguistic content remains an open question (Ocampo et al. 2023).

In addition, any action must carefully consider the dangers of transferring regulation and enforcement to private companies. (Stark et al. 2020) argue that over-filtering content is just as dangerous as letting bad content spread. Indeed, deletion and filtering may deviate from initial purpose to over-censorship of content if it becomes safer for the platforms to do so than take a risk of being sued. Furthermore, assessing the trustworthiness of information raises multiple ethical and political concerns: Who decides what is true or false? According to which criteria? Under whose control?

Secondly, to mitigate the effect of false information, multiple post hoc measures rely on the impact of additional corrective content. For (Vraga and Bode 2020), pointing to a coherent alternative explanation, with references to expert and highly credible factual sources, remains a solid starting point. The authors describe the strategy of “observational correction” leveraging the fact that users who witness the correction of a misinformation item, but have not directly engaged with that item, are less affected by cognitive dissonance and are therefore more amenable to correction. This is consistent with the findings of (Bode and Vraga 2015) who suggest that exposing users to related stories that correct a post that contained misinformation will significantly reduce misperceptions. The impact of the correction can be further reinforced by explicitly pointing to the demographic similarity between the user and the authors of opposing content (Garrett, Nisbet, and Lynch 2013), which taps into the homophily effect⁸. In other words, we are more likely to accept the correction when it comes from someone who is socially close to us, e.g. having the same professional activity or background. (Vraga and Bode 2020) also suggest to multiply correction actions for each targeted content to reinforce the effect.

6 Reclaiming Our Attention

The methods presented in the previous section all have one thing in common: they deal with the problems in a post hoc manner, that is, after these problems have occurred, with all the limitations that come with this “coming after”. To go further however, we need to figure out measures, may they be legal, economical or technical, capable of preventing the attention from being looted in the first place. More

importantly, we need to consider this reflection not only from the perspective of regulating the attention consumers (the platforms and multiple intermediaries), but also from the perspective of the producers (the end-users) who want to reclaim their attention, especially in times when our attention is needed on a number of urgent matters. This involves actively preventing recommendation systems from finding ways to exploit our inner limitations and manipulate us through sometime ancient and deeply embedded structures of our brain (e.g. our striatum) (Bohler 2019).

Below we formulate a set of propositions stemming from the observations and findings reported in the previous sections. We organize them around the challenges that they address, together with suggestions made by other authors from multiple disciplines. Finally we extract from them a set of empirical principles that could be used to drive further works on good practices.

The Carrot and the Stick

Taking the example of false information, (Vraga and Bode 2020) insist on the fact that a posteriori corrections are not sufficient and must happen as early as possible, that is, before misperceptions are entrenched. Besides, avoiding the algorithmic amplification effect of such information by recommendation systems requires to mitigate the popularity effect before it happens (Fernández, Bellogín, and Cantador 2021). But if online social media are required by law to combat false information, they have conflicting incentive to do it, not to say no incentive at all. Indeed, as we described in section 3, false information largely relies on negative emotions to capture users’ attention. As such, they are very effective in fostering user engagement which is what online social media strive to obtain. Consequently, from an economic point of view, it is counter-productive for them to prevent the spread of false information. More generally, it is counter-productive for platforms to mitigate the popularity effect, mitigate the impact of negative emotions, or reduce filter bubbles and the subsequent polarization of opinions.

The authors of (Newman 2019) suggest to rethink existing trade regulation laws such as antitrust and fair competition laws under the new realm of attention markets. They propose to enforce taxes on attention consumption to “disincentivize attention intermediaries from vacuuming up as much attention as possible”, for instance by restraining the amount of advertisements that can be shown to a user, or reducing the deductibility of advertising expenditures from the companies’ revenues to alleviate their taxes. They also propose to regulate the attention costs that can be charged, with the idea that if attention becomes less lucrative then financial resources will be redirected towards more lucrative markets, thus reducing the amount of attention being captured and traded. Such regulation could take the form of a tax on the daily attention time being caught beyond a certain threshold. Another tax could be enforced on the individual and collective harms caused by attention capturing techniques, which first requires to explicit those harms like it has been done for dark patterns (Santos, Morozovaite, and De Conca 2024) and more precisely to qualify (Beckers, Chockler, and Halpern 2022) and quantity (Beckers, Chock-

⁸Homophily: the tendency to associate with similar others.

ler, and Halpern 2023) the harms done.

In other words, things would not change without strong incentives on one side, and disincentives on the other. We can summarize this in the following general principle:

📌 *principle of the right incentive*

Governance bodies should leverage legal and economic means to drive platforms' practices towards desirable behaviours, while penalizing undesired behaviours.

Usage Regulation

Some of the measures meant to regulate the attention market lie in the way the services are consumed. As has already been done in some countries, laws could be voted to limit the daily time spent by users on certain services, especially among the youngest (Kantowitz 2021; Carroll and correspondent 2023). Another simple measure applies to video streaming platforms, that consists in imposing few-second pauses between videos. This apparently naive technique may actually shatter the infinite feed trap by giving users the short amount of time they need to realize that they have been in an attention tunnel for a while, and that they want to "reclaim" their attention. This can be generalized by formulating the following principle:

📌 *principle of supported due diligence*

All means should be provided to foster and update the due diligence of users. In particular they should always be given and made aware of the options to escape the systems' loops, processes and goals.

Policy makers could also tackle the problem of attention fragmentation entailed by the multiple, often invasive, notifications that smartphone applications raise. Whenever a notification occurs, users are tempted to interrupt their current activity, check the reason of the notification, possibly react to it, before eventually returning to their activity. It has been shown that switching our attention between tasks or contexts has a cost: it is time-consuming and creates a more error prone context (Jersild 1927; Monsell 2003; Rogers and Monsell 1995). Furthermore it has even been shown that the mere presence of such devices, although turned off, impairs our cognitive capacity (Ward et al. 2017). In a mindset similar to the European General Data Protection Regulation (GDPR)⁹, which imposes the consent of users for the use of cookies, the EU Parliament has recently adopted a resolution on addictive design to put forward "a digital 'right not to be disturbed' to empower consumers by turning all attention-seeking features off by design" (The European Parliament 2023). This could lead to laws imposing that smartphone applications obtain users' explicit and informed consent for the notifications that they raise, and deactivate them by default ("opt-in only"). Hence the following principle:

⁹General Data Protection Regulation (GDPR)
<https://gdpr-info.eu/>

📌 *principle of opt-in by default*

Recommendation and notification services should be turned off by default and only turned on on demand and after informed consent and preference setting

This could be complemented by more punitive measures, as proposed in (Mudde and Kaltwasser 2017), for instance by demonetizing and forbidding collaboration with platforms that do not follow the rules.

Content Recommendation Monitoring

The echo-chamber effect of recommendation algorithms is at the root of multiple examples of polarization and radicalization. It could be mitigated by imposing a certain share of non-recommended content, content that is outside of the user's interests, or content that originates from users they are not acquainted with. In this respect, some approaches lie in the second category proposed by (Fernández, Bellogín, and Cantador 2021), i.e. modifications "within the recommendation" system. For instance, the same authors suggest using clustering approaches to assemble the contacts of the user according to different levels of similarity with the user, and leverage these groups to increase the diversity of recommendations while maintaining a certain coherence and similarity. (Fabbri et al. 2022) propose a method to come up with relevant recommendations while reducing the likelihood of enticing the user towards radicalization pathways. Also, to counter the misuse and abuse of anger, indignation or fear, which are often associated with false information, platforms could be required to carry out sentiment analysis on every content in order to keep the amount of recommendations associated to negative emotions below a given threshold.

📌 *principle of balanced recommendations*

Recommendation-based platforms should prevent the over specialization of recommendations w.r.t. all features and should support monitoring and prevention of the formation of bubbles of any type (opinion, source, emotion, etc.).




Moreover, there exists an asymmetry of visibility between the viral spreading of an information that was proven to be false or misleading, and the denial or rectification of that information. The denial of a false information usually puts forward a pondered, nuanced position that appeals to reasoning and facts (*logos*) over emotion (*pathos*). Hence, it does not trigger an emotional response compared to the one generated by the false information in the first place, and it is therefore silently downgraded by recommendation algorithms. This is commonly summarized by the so-called Bradolini's law which states that "the amount of energy needed to refute false information is an order of magnitude bigger than that needed to produce it." As a result, users who propagate false information often never get to know about their mistake. (Venturini 2019) insist on the fact that it is critical to jointly address content-checking and digital virality. Thus, to counterbalance this visibility asymmetry, social media could be required to impose on the denial/rectification of a false

information a visibility equivalent to that of the initial information, for instance, by ensuring that the population who was exposed to the false information be exposed to the denial too. A warning could also be presented to users who propagated this false information in order to increase their awareness. Of course, this type of measure could be coupled with other post recommendation measures such as strategies involving the observational correction or demographic similarity presented in section 5. More generally, one has to figure out how we can use recommendation systems to recommend counter measures, *i.e.* we could train a recommendation system to learn the most relevant content and the most impactful entries in the acquaintance network to inject a counter measure.

📌 principle of balanced visibility

Recommendation-based platforms should ensure that preventive and corrective measures have a visibility at least equal to the visibility of the problems being prevented or corrected.

Affordances and Interaction Design

As discussed in section 3, the frictionless affordances of platforms are optimized towards extremely brief and basic exchanges, to favor their quantity over their quality, leaving no room for nuance, pondering, doubt nor substantiated reasoning. Interfaces could be redesigned to facilitate non-binary reactions, starting with a range of nuanced emotions. Rather than implementing deceptive dark patterns, they could restore “desirable frictions” (Broadbent et al. 2024) to prevent compulsive, consciousness behaviors, and rely on nudges to gently drive users towards critical thinking, and by valuing/rewarding this kind of behavior. (Abreu and França 2021) recommend engaging users in the validation of content before sharing it, both manually and with automated analysis methods on content and context. For instance, X (formerly Twitter) asks confirmation before retweeting the link to an article that the user did not click. Similarly, interfaces could encourage users to comment on content instead of merely clicking ,  or , and they could question a user about whether they really want to share or support a content associated to strong negative emotions or for which a counter-measure was triggered.

📌 principle of benevolent interaction design

Affordances and interactions should be designed and evaluated with the well-being of end-users in mind first.

Societal Impact and Educational Mission

We, as a society, could decide that large online social media, because of their influence on the society, public opinion, public health and economy, can no longer be considered as sheer private companies regulated by markets law only. Instead, they could be seen as digital commons and be assigned a specific status that would endow them with a societal mission including, for instance, an educational purpose. As an example, they could instruct users in detecting false or misleading information, they could promote content meant

to increase awareness w.r.t. attention mechanisms and cognitive biases, foster assertiveness, critical thinking and “distill” concepts of the scientific method, etc. On the same page, authors of (Abreu and França 2021) insist on the need for civic education, and (O’neil 2017) recommend integrating democratic values into the algorithms that impact our lives, especially the ones participating in an algorithmic governance (e.g. platform used for debates, for information, for legal actions, education). States and civil societies could favor the entrance of new platforms that, instead of optimizing user engagement and attention catching, would favor democratic debate in the digital space by supporting free constructive debate and the collective development of solutions to major societal issues. An illustration of this approach is that of Polis, an open-source social network meant to support large-scale public debate (Small et al. 2021). Its recommendation algorithm is trained to promote consensus. During experiments, particularly in Taiwan, Polis has “shown itself capable of building shared understanding, disincentivizing counterproductive behavior (trolling), and cultivating points of consensus”.

📌 principle of digital commons preservation

When a digital service, platform or resource reaches the potential of having a world-wide impact on human societies, it must be assigned the status of digital common and must be subjected to preservation rules and policies.

Feedback and Transparency Enforcement

One of the pitfalls we identified is the fact that users are not aware of being caught by recommendation loops. Furthermore, it is established that online sharing of fake news increases with social media fatigue (Talwar et al. 2019). Therefore, approaches such as *quantified self* and *lifelog* could be specialized to the case of recommendation-based platforms, in order to foster awareness and introspection. Some online services already provide their users with usage metrics with respect to the total time spent on the platform, or reminders to make a break at regular intervals. This can be seen as a shy improvement towards compliance with the European Digital Services Act (Union 2022), that took effect in August 2023, however such options remain deactivated by default, and usually accessible from impractical sub-menus. Self-tracking tools could provide users with a much larger set of metrics, indicators and feedback such as the total exposure to negative news, the diversity of the recommendations they are shown, and make them aware of low-diversity risks. For instance the fact that “90% of the content one sees come from 10% of one’s contacts or are on the same topic” may indicate that one is experiencing a filter bubble. In other words, platforms must set up the conditions for individual and collective reflexivity, that is, the ability of users and communities to be self-aware of their usage and engagement with the system. Hence the following principle:

📌 principle of continuous reflexivity

Users must be provided a continuously updated feedback on their usage of the system and on themselves to support their reflexivity and maintain an up-to-date informed consent.

This reflexivity requires not only usage metrics but also transparency about the platform's purposes. Among other measures, the European Digital Services Act requires that platforms set up mechanisms to explain the reasons that led to recommending a certain content, and to offer users an alternative recommendation not based on profiling. Such measures are especially crucial when coupling AI and the Web since we need to set transparency and explanation as a prerequisite to any approach, to ensure the awareness and informed consent of, potentially, billions of users (Berendt et al. 2021).

📌 principle of full user awareness

Users must be made aware of all the features and purposes leading to a recommendation, before and when it is provided.

Let us add that, for the regulations proposed above to be enforced effectively, one cannot rely solely on the goodwill of online platforms. But given the scale of these platforms, public actors have neither the information nor the technical possibilities to evaluate the application and effectiveness of the regulations (Zolynski, Le Roy, and Levin 2019). Hence the following principle:

📌 principle of observability

Online platforms must provide institutions, civil society and researchers with the legal and technical instruments allowing them to carry out active control and verification of the application and effectiveness of regulations.

To the very least, this requires open and free access to platforms data by researchers, as opposed to the recent decision of X to deny them free API access.

Build on Existing Practices

Finally, and although it may seem obvious, one rule is worth remembering: to review and take inspiration from existing best practices in other domains. In most jurisdictions there exist advertising laws to protect consumers, ensure they remain able to make informed decisions, and more generally to maintain a level playing field¹⁰ between all players. Most countries regulate advertising through legislations that target different forms of false, misleading or deceptive advertising contents and claims, and forbid a whole range of practices (unsubstantiated comparison, forged testimonial, puffery, misleading packaging/label, unsolicited commercial messages, alleged contests and sweepstakes, etc.). The work and literature on regulating advertising should be

¹⁰Metaphor denoting the fact that, in business, all players compete fairly, i.e. they all play by the same set of rules. https://en.wikipedia.org/wiki/Level_playing_field

reviewed and built-upon in regulating the attention market at large. This topic is also close to that of clickbaits that are recommended links designed to attract attention and to entice users to follow them while being typically deceptive, sensationalized, or otherwise misleading. Clickbaits are not just teasers but headlines with an element of dishonesty, “using enticements that do not accurately reflect the content being delivered”¹¹. As far as we know, there is no regulation of clickbait practices on the Web, although some of these techniques bear similarities with the misleading or deceptive advertisement practices that we just mentioned and that, on the contrary, are regulated.

📌 principle of best practice transfer

Methods and tools used to regulate similar situations in relevant domains should be surveyed, benchmarked and systematically considered as input to a Web and AI governance.

To give another example coming from a completely different angle, we know that parenting practices in terms of TV viewing have an impact on the behaviour of young watchers (Barradas et al. 2007). Again, approaches and good practices in this domain, and more generally in educating and parenting in the digital media age (Coyne et al. 2017), must be considered in the case of “Web viewing” in general and when addressing the problem of attention capturing in particular.

More generally speaking, we need to put in place a governance bodies, starting with the Web and AI, that are prepared to tackle new problematic practices and regulate them, as is done in other areas of activity. And we also need to keep a constant watch on these other areas, if only to draw inspiration from the initiatives and feedback they have on similar issues. Taking the example of the video game industry, there is evidence of a relationship between “loot box”¹² spending and gambling addiction (Zendle and Cairns 2018), and that a loot box is psychologically akin to gambling (Drummond and Sauer 2018) and can result in addictive behaviors and endangered players. The way to study and address that unwanted exploitation of our behaviors is inspirational for other problematic practices on the Web such as those we surveyed.

7 Thank you for... your Attention

AI is domain-independent. It is being applied in all of our daily life areas of interest: information, business, money, politics, employment, sports, games, sex etc. And the worldwide deployment of these techniques, partly due to its coupling with the Web, could have detrimental consequences in all these areas alike, unless properly regulated. This is a commonplace observation but it is the reason why, to prevent such detrimental effects, an ethical AI approach to AI governance must be multidisciplinary and interdisciplinary.

¹¹Definition adapted from <https://en.wikipedia.org/wiki/Clickbait>

¹²“loot boxes” are video game items with randomized contents that can be paid for with real-world money.

With this mindset, this paper brought together conclusions from more than 70 articles and books from different disciplines (psychology, sociology, neuroscience, politics, legal domain, computer science, education, etc.) to analyze and call for actions against the current practices competing to capture our attention in several “Web Wild West corners”. The problem is both critical and complex, and authors of (Kitchens, Johnson, and Gray 2020) defend the need for a “nuanced multidimensional view of how social media use may shape information consumption” and they urge us to consider “the complex variety in social media platforms [and the] considerable variation in observed impacts among them”. In (Fernández, Bellogín, and Cantador 2021) authors add that “This research requires to navigate the careful tension between privacy, security, economic interests, censorship and cultural differences, and requires to be addressed from multiple disciplines that can assess not only the technological aspect, but also the individual and the social one (...) There is ample room for investigation (...), opening a novel, exciting and interdisciplinary line of research.”

At the same time, the problem is getting worse with every technological innovation. The pervasiveness of smartphones in our lives has further reinforced the effectiveness of these techniques that can now grab our attention at every moment of the day, and in particular these moments that were previously those of boredom, waiting, daydreaming or intellectual strolling. As we pointed out in section 4, these moments are known to be necessary to spur imagination and creativity. In the continuation of smartphones, smart objects and the resulting internet of things and Web of things will only make things worse.

Recommendation systems that learn to predict us effectively learn to manipulate us, and to be predictable is to lose freedom. Everyday we fuel the predictors in exchange for immediate satisfaction and instant pleasure, this amounts to continually mortgaging our freedom. Besides, these systems that compete for our attention end up pressuring us to consume and to react more and more quickly to their recommendations. And, as we know, acceleration is a form of alienation (Rosa and Chaumont 2017).

In another context and to address our own human limitations, (Bronner 2021) recommended to find ways to increase our overall level of consciousness and reclaim the power of long-term reflection. Our leaders and role models¹³ struggle to embody the values of patience, conscience and moderation (Bronner 2021), but our computer systems rarely drive us in that direction. On the contrary, current AI applications are pushing us not to use our conscience, but to play their automation game. Yet there is no reason for these systems to live in our mind rent free and it is urgent to redesign them so they regularly push us to take a step back, to be more conscious of what we are doing, viewing, saying, spreading, etc. The challenge is to (re)take and (re)give time for awareness, attention and reflection: we need to (re)claim that source of freedom. And for this, we proposed a non-exhaustive first set of principles to (re)design applications and inscribe in them

¹³A role model is a person whom others look at as an example to be imitated.

a set of agreed-upon values.

References

- Abreu, C.; and França, L. A. 2021. Algorithm-driven populism: An introduction [Populizm oparty na algorytmach. Wprowadzenie.]. *Archives of Criminology [Archiwum Kryminologii]*.
- Acerbi, A. 2019. Cognitive attraction and online misinformation. *Palgrave Communications*, 5(1): 1–7.
- Barradas, D. T.; Fulton, J. E.; Blanck, H. M.; and Huhman, M. 2007. Parental Influences on Youth Television Viewing. *The Journal of Pediatrics*, 151(4): 369–373.e4.
- Basile, V.; Cabrio, E.; Villata, S.; Frasson, C.; and Gandon, F. 2016. A Pragma-Semantic Analysis of the Emotion/Sentiment Relation in Debates. In Lieto, A.; Bhatt, M.; Oltramari, A.; and Vernon, D., eds., *Proceedings of the 4th International Workshop on Artificial Intelligence and Cognition co-located with the Joint Multi-Conference on Human-Level Artificial Intelligence (HLAI 2016), New York City, NY, USA, July 16-17, 2016*, volume 1895 of *CEUR Workshop Proceedings*, 117–123. CEUR-WS.org.
- Beckers, S.; Chockler, H.; and Halpern, J. 2022. A causal analysis of harm. *Advances in Neural Information Processing Systems*, 35: 2365–2376.
- Beckers, S.; Chockler, H.; and Halpern, J. Y. 2023. Quantifying harm. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI ’23*. ISBN 978-1-956792-03-4.
- Benlamine, M. S.; Chaouachi, M.; Villata, S.; Cabrio, E.; Frasson, C.; and Gandon, F. 2015. Emotions in Argumentation: an Empirical Evaluation. In Yang, Q.; and Wooldridge, M. J., eds., *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, 156–163. AAAI Press.
- Berendt, B.; Gandon, F.; Halford, S.; Hall, W.; Hendler, J.; Kinder-Kurlanda, K. E.; Ntoutsis, E.; and Staab, S. 2021. Web Futures: Inclusive, Intelligent, Sustainable The 2020 Manifesto for Web Science. *Dagstuhl Manifestos*.
- Board, E. D. P. 2022. Guidelines 03/2022 on deceptive design patterns in social media platform interfaces: how to recognise and avoid them - v2/0. Technical report.
- Bode, L.; and Vraga, E. K. 2015. That Was Wrong: The Correction of Misinformation Through Related Stories Functionality in Social Media. *Journal of Communication*, 65(6): 619–638.
- Bohler, S. 2019. *Le Bug humain: Pourquoi notre cerveau nous pousse à détruire la planète et comment l’en empêcher*. Robert Laffont.
- Broadbent, S.; Forestier, F.; Khamassi, M.; and Zolynski, C. 2024. *Pour une nouvelle culture de l’attention*. Odile Jacob.
- Bronner, G. 2021. *Apocalypse cognitive*. Presses Universitaires de France. ISBN 978-2-13-073304-1.
- Carroll, R.; and correspondent, R. C. I. 2023. ‘Much easier to say no’: Irish town unites in smartphone ban for young children. *The Guardian*.

- Chuai, Y.; and Zhao, J. 2022. Anger can make fake news viral online. *Frontiers in Physics*, 10: 970174.
- Cialdini, R. B. 2008. *Influence: Science and practice*. Pearson Education.
- Corbu, N.; Bârgăoanu, A.; Durach, F.; and Udrea, G. 2021. Fake News Going Viral: The Mediating Effect Of Negative Emotions. *Media Literacy and Academic Research*, 4(2): 58–87.
- Courbet, D.; Fourquet-Courbet, M.-P.; Kazan, R.; and Intartaglia, J. 2014. The Long-Term Effects of E-Advertising: The Influence of Internet Pop-ups Viewed at a Low Level of Attention in Implicit Memory. *Journal of Computer-Mediated Communication*, 19(2): 274–293.
- Coyne, S. M.; Radesky, J.; Collier, K. M.; Gentile, D. A.; Linder, J. R.; Nathanson, A. I.; Rasmussen, E. E.; Reich, S. M.; and Rogers, J. 2017. Parenting and Digital Media. *Pediatrics*, 140(Supplement 2): S112–S116.
- Crockett, M. J. 2017. Moral outrage in the digital age. *Nature Human Behaviour*, 1(11): 769–771.
- DeVito, M. A. 2017. From editors to algorithms: A values-based approach to understanding story selection in the Facebook news feed. *Digital journalism*, 5(6): 753–773.
- Drummond, A.; and Sauer, J. D. 2018. Video game loot boxes are psychologically akin to gambling. *Nature human behaviour*, 2(8): 530–532.
- Fabbri, F.; Wang, Y.; Bonchi, F.; Castillo, C.; and Mathioudakis, M. 2022. Rewiring What-to-Watch-Next Recommendations to Reduce Radicalization Pathways. In *Proceedings of the ACM Web Conference 2022, WWW '22*, 2719–2728. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-9096-5.
- Fan, R.; Zhao, J.; Chen, Y.; and Xu, K. 2014. Anger Is More Influential than Joy: Sentiment Correlation in Weibo. *PLOS ONE*, 9(10): e110184.
- Fernández, M.; Bellogín, A.; and Cantador, I. 2021. Analysing the Effect of Recommendation Algorithms on the Amplification of Misinformation. ArXiv:2103.14748 [cs].
- Fortunato, P.; and Pecoraro, M. 2022. Social media, education, and the rise of populist Euroscepticism. *Humanities and Social Sciences Communications*, 9(1): 1–13.
- Funke, D.; and Flamini, D. 2023. A guide to anti-misinformation actions around the world.
- Garrett, R. K.; Nisbet, E. C.; and Lynch, E. K. 2013. Undermining the corrective effects of media-based political fact checking? The role of contextual cues and naïve theory. *Journal of Communication*, 63(4): 617–637.
- Gray, C. M.; Sanchez Chamorro, L.; Obi, I.; and Duane, J.-N. 2023. Mapping the Landscape of Dark Patterns Scholarship: A Systematic Literature Review. In *Designing Interactive Systems Conference*, 188–193. Pittsburgh PA USA: ACM. ISBN 978-1-4503-9898-5.
- Harris, T. 2019. How Technology is Hijacking Your Mind — from a Former Insider.
- Hefti, A.; and Heinke, S. 2015. On the economics of superabundant information and scarce attention. *Æconomia. History, Methodology, Philosophy*, (5-1): 37–76. Number: 5-1 Publisher: Association Œconomia.
- Hendricks, V. F.; and Vestergaard, M. 2019. The Attention Economy. In Hendricks, V. F.; and Vestergaard, M., eds., *Reality Lost: Markets of Attention, Misinformation and Manipulation*, 1–17. Cham: Springer International Publishing. ISBN 978-3-030-00813-0.
- Intelligence, G. T. 2022. Social media, algorithms, and populism.
- Jersild, A. T. 1927. *Mental set and shift*. 89. Columbia university.
- Kantrowitz, A. 2021. 5 Ways China is Mandating Social Media Changes.
- Keib, K.; Espina, C.; Lee, Y.-I.; Wojdowski, B. W.; Choi, D.; and Bang, H. 2018. Picture This: The Influence of Emotionally Valenced Images, On Attention, Selection, and Sharing of Social Media News. *Media Psychology*, 21(2): 202–221.
- Kitchens, B.; Johnson, S. L.; and Gray, P. 2020. Understanding Echo Chambers and Filter Bubbles: The Impact of Social Media on Diversification and Partisan Shifts in News Consumption. *MIS Quarterly*, 44(4): 1619–1649.
- Kohout, S.; Kruikeimeier, S.; and Bakker, B. N. 2023. May I have your Attention, please? An eye tracking study on emotional social media comments. *Computers in Human Behavior*, 139: 107495.
- Kramer, A. D. 2012. The spread of emotion via Facebook. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '12*, 767–770. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-1015-4.
- Lewis, P. 2017. 'Our minds can be hijacked': the tech insiders who fear a smartphone dystopia. *The Guardian*.
- Martel, C.; Pennycook, G.; and Rand, D. G. 2020. Reliance on emotion promotes belief in fake news. *Cognitive Research: Principles and Implications*, 5(1): 47.
- Mascarell, A. B. 2020. A bibliometric analysis of utopian literature. *ES Review. Spanish Journal of English Studies*, (41): 77–103.
- Monsell, S. 2003. Task switching. *Trends in cognitive sciences*, 7(3): 134–140.
- Mudde, C.; and Kaltwasser, C. R. 2017. *Populism: A very short introduction*. Oxford University Press.
- Newman, J. M. 2019. Regulating Attention Markets. *University of Miami Legal Studies Research Paper*.
- Ocampo, N.; Sviridova, E.; Cabrio, E.; and Villata, S. 2023. An In-depth Analysis of Implicit and Subtle Hate Speech Messages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, 1997–2013*. Dubrovnik, Croatia: Association for Computational Linguistics.
- O'neil, C. 2017. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.
- Pariser, E. 2011. *The filter bubble: How the new personalized web is changing what we read and how we think*. Penguin.

- Parkin, S. 2018. The YouTube stars heading for burnout: 'The most fun job imaginable became deeply bleak'. *The Guardian*.
- Patino, B. 2024. L'émergence du caudillo numérique. *The Tocqueville Review*. Publisher: University of Toronto Press.
- Ra, C. K.; Cho, J.; Stone, M. D.; De La Cerda, J.; Golden-son, N. I.; Moroney, E.; Tung, I.; Lee, S. S.; and Leventhal, A. M. 2018. Association of Digital Media Use With Subsequent Symptoms of Attention-Deficit/Hyperactivity Disorder Among Adolescents. *JAMA*, 320(3): 255–263.
- Richards, B. 2007. *Emotional governance: Politics, media and terror*. Springer.
- Richards, J. M.; and Gross, J. J. 2000. Emotion regulation and memory: the cognitive costs of keeping one's cool. *Journal of personality and social psychology*, 79(3): 410.
- Robertson, C. E.; Pröllochs, N.; Schwarzenegger, K.; Pärnamets, P.; Van Bavel, J. J.; and Feuerriegel, S. 2023. Negativity drives online news consumption. *Nature Human Behaviour*, 7(5): 812–822.
- Rocha, Y. M.; de Moura, G. A.; Desidério, G. A.; de Oliveira, C. H.; Lourenço, F. D.; and de Figueiredo Nico-lete, L. D. 2023. The impact of fake news on social media and its influence on health during the COVID-19 pandemic: a systematic review. *Journal of Public Health*, 31(7): 1007–1016.
- Rogers, R. D.; and Monsell, S. 1995. Costs of a predictable switch between simple cognitive tasks. *Journal of experi-mental psychology: General*, 124(2): 207.
- Rosa, H.; and Chaumont, T. 2017. *Aliénation et accélération: vers une théorie critique de la modernité tar-dive*. La découverte.
- Rouvroy, A.; and Berns, T. 2013. Gouvernamentalité algo-rithmique et perspectives d'émancipation. *Réseaux*, 177(1): 163–196.
- Santos, C.; Morozovaite, V.; and De Conca, S. 2024. No harm no foul: how harms caused by dark patterns are con-ceptualised and tackled under EU data protection, consumer and competition laws.
- Sasahara, K.; Chen, W.; Peng, H.; Ciampaglia, G. L.; Flam-mini, A.; and Menczer, F. 2021. Social Influence and Un-following Accelerate the Emergence of Echo Chambers. *Journal of Computational Social Science*, 4(1): 381–402. ArXiv:1905.03919 [physics].
- Schonfeld, A.; McNeil, D.; Toyoshima, T.; and Binder, R. 2023. Cyberbullying and Adolescent Suicide. *Journal of the American Academy of Psychiatry and the Law Online*. Publisher: Journal of the American Academy of Psychiatry and the Law Online Section: Analysis and Commentary.
- Shah, N. 2022. Making TikTok Videos Leaves Musicians Feeling Burnout. *Wall Street Journal*.
- Sharma, K.; Qian, F.; Jiang, H.; Ruchansky, N.; Zhang, M.; and Liu, Y. 2019. Combating Fake News: A Survey on Iden-tification and Mitigation Techniques. ArXiv:1901.06437 [cs, stat].
- Siegrist, M.; and Cvetkovich, G. 2001. Better negative than positive? Evidence of a bias for negative information about possible health dangers. *Risk Analysis: An Official Publica-tion of the Society for Risk Analysis*, 21(1): 199–206.
- Small, C.; Jorkegren, M.; Erkkilä, T.; Shaw, L.; and MEGILL, C. 2021. Polis: Scaling Deliberation by Mapping High Dimensional Opinion Spaces. *RECERCA. Revista de Pensament i Anàlisi*.
- Smith, D. R. 2018. Attention, attention: your most valu-able scientific assets are under attack. *EMBO reports*, 19(3): e45684. Publisher: John Wiley & Sons, Ltd.
- Soroka, S.; Fournier, P.; and Nir, L. 2019. Cross-national ev-idence of a negativity bias in psychophysiological reactions to news. *Proceedings of the National Academy of Sciences*, 116(38): 18888–18892.
- Stark, B.; Stegmann, D.; Magin, M.; and Jürgens, P. 2020. Are algorithms a threat to democracy? The rise of interme-diaries: A challenge for public discourse. *Algorithm Watch*, 26.
- Suler, J. 2004. The Online Disinhibition Effect. *CyberPsy-chology & Behavior*, 7: 321–326.
- Talwar, S.; Dhir, A.; Kaur, P.; Zafar, N.; and Alrasheedy, M. 2019. Why do people share fake news? Associations be-tween the dark side of social media use and fake news shar-ing behavior. *Journal of Retailing and Consumer Services*, 51: 72–82.
- The European Parliament. 2023. European Parliament res-olution of 12 December 2023 on addictive design of online services and consumer protection in the EU single market (2023/2043(INI)).
- Times, T. E. 2017. No time to kill: How smartphone is push-ing chewing gum out of fashion.
- Union, E. 2022. Regulation (EU) 2022/2065 of the Euro-pean Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act). *Official Journal of the European Union*.
- Venturini, T. 2019. From fake to junk news: The data politics of online virality. In *Data politics*, 123–144. Routledge.
- Villata, S.; Benlamine, M. S.; Cabrio, E.; Frasson, C.; and Gandon, F. 2018. Assessing Persuasion in Argumentation through Emotions and Mental States. In Brawner, K.; and Rus, V., eds., *Proceedings of the Thirty-First International Florida Artificial Intelligence Research Society Conference, FLAIRS 2018, Melbourne, Florida, USA. May 21-23 2018*, 134–139. AAAI Press.
- Vosoughi, S.; Roy, D.; and Aral, S. 2018. The spread of true and false news online. *Science*, 359(6380): 1146–1151.
- Vraga, E. K.; and Bode, L. 2020. Correction as a solution for health misinformation on social media. *American Journal of Public Health*, 110(S3): S278–S280.
- Ward, A. F.; Duke, K.; Gneezy, A.; and Bos, M. W. 2017. Brain drain: The mere presence of one's own smartphone reduces available cognitive capacity. *Journal of the Associ-ation for Consumer Research*, 2(2): 140–154.
- Waszak, P. M.; Kasprzycka-Waszak, W.; and Kubanek, A. 2018. The spread of medical fake news in social media –

The pilot quantitative study. *Health Policy and Technology*, 7(2): 115–118.

Whateley, D. 2023. TikTok's music influence is 'exhausting' artists and marketers alike as the industry grapples with the pressure to go viral.

Whittaker, J.; Looney, S.; Reed, A.; and Votta, F. 2021. Recommender systems and the amplification of extremist content. *Internet Policy Review*, 10(2).

Zaeem, R. N.; Li, C.; and Barber, K. S. 2020. On Sentiment of Online Fake News. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 760–767. ISSN: 2473-991X.

Zendle, D.; and Cairns, P. 2018. Video game loot boxes are linked to problem gambling: Results of a large-scale survey. *PloS one*, 13(11): e0206767.

Zolynski, C.; Le Roy, M.; and Levin, F. 2019. L'économie de l'attention saisie par le droit. *Dalloz IP/IT : droit de la propriété intellectuelle et du numérique*, (11): 614.

Zomorodi, M. 2017. *Bored and brilliant: How time spent doing nothing changes everything*. Pan Macmillan.

Zuboff, S. 2019. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. Profile Books. ISBN 9781781256855.