LANDESANSTALT FÜR MEDIEN NRW
Der Meinungsfreiheit verpflichtet.

DISINFORMATION
RISKS, REGULATORY GAPS AND ADEQUATE COUNTERMEASURES

Expert Opinion Commissioned by the Landesanstalt für Medien NRW

Stephan Dreyer, Elena Stanciu, Keno Christopher Potthast, Wolfgang Schulz

Carried out by

LEIBNIZ-INSTITUT
FÜR MEDIENFORSCHUNG
HANS-BREDOW-INSTITUT

CONTENTS

1. BACKGROUND AND AIM OF THE STUDY

Disinformation is by no means a novel phenomenon. However, with the transformation of public communication and its forums, the forms of its appearance, its terms used in discourse, its reach or visibility and the contexts and types of effects of disinformation change as well. Possibilities to produce professional content that gets distributed very quickly (by either humans or automated networks of actors) and that shows high degrees of user activation and interaction have boosted disinformation as a hot societal topic. The massive, sometimes viral distribution of relevant depictions and statements has made it evident that not only the content of such messages itself, but especially the combination of content and its vast reach bears individually and societally relevant risk potentials.[1] This necessitates the engagement of all societal groups with the topic of disinformation. Hence, disinformation-related discussions can currently be observed in scientific, political, societal and legal discourses.

However, many of these discourses tend to be alarmist: Disinformation destroys either the democracy, its foundations, society, social cohesion, public peace, shared realities – or all of them together. Calls for a Strong State that keeps disinformation within bounds and (re-)stabilizes society are resounding clearly while the classic *Böckenförde* dilemma appears to be partially forgotten: "The liberal, secularized state lives by prerequisites which it cannot guarantee itself."[2] Statutory prohibitions and legal countermeasures are being passionately discussed.[3] Given the large number of directly or indirectly affected stakeholders, the societal debate appears to be a disinformation ostracizing cacophony. Depending on the counting method, eight to twelve regulatory developments and policy initiatives are underway on the European level[4] that (partly) focus on disinformation. Compared to that, the political and legislative agenda in Germany appears to be comparatively quiet.

The use of statutory measures to control disinformation phenomena is not trivial. At the heart of the matter these approaches affect the statements of individuals, which are usually not only permissible but also covered by the freedom of expression. The limits of what can be said, which arise primarily in view of hate speech cannot simply be transferred to disinformation. Instead, this requires a systematic examination as to the possibilities and limitations of legal intervention in this area.

Such a systematic analysis of forms of disinformation and their risk potentials against the background of the existing legal framework has not been conducted so far and therefore represents a relevant approach. It was conducted on behalf of the Landesanstalt für Medien NRW. Based on the identified gaps in legal protection, the study develops potential countermeasures and describes necessary conditions for their effectiveness. The core question is: **What are the risk potentials of disinformation and which governance measures can be chosen to adequately respond to these risks?**

This key question is being approached in three steps (see Figure 1): In a first step (Chapter 2), types of disinformation that appear as differentiated in scientific and policy discussions are being identified, their respective definitions and notions are summarized and they are investigated with regard to their risk potential. The aim is to show the extent of the relevant phenomena as well as to distinguish them from other phenomena and terminologies. While doing so, the indicators use to differentiate are also assessed in terms of their usability for legal and/or regulatory use. Moreover, the state of research with regard to the detrimental effects of disinformation for legally protected rights and societal goals is taken into account; to date, only scattered knowledge of the effects of disinformation on individual recipients is available. This contrasts with the often implied suggestions regarding disinformation that current regulatory demands are based on. Where empiric evidence is available, the study points out presumed effects and their respective risk potential.

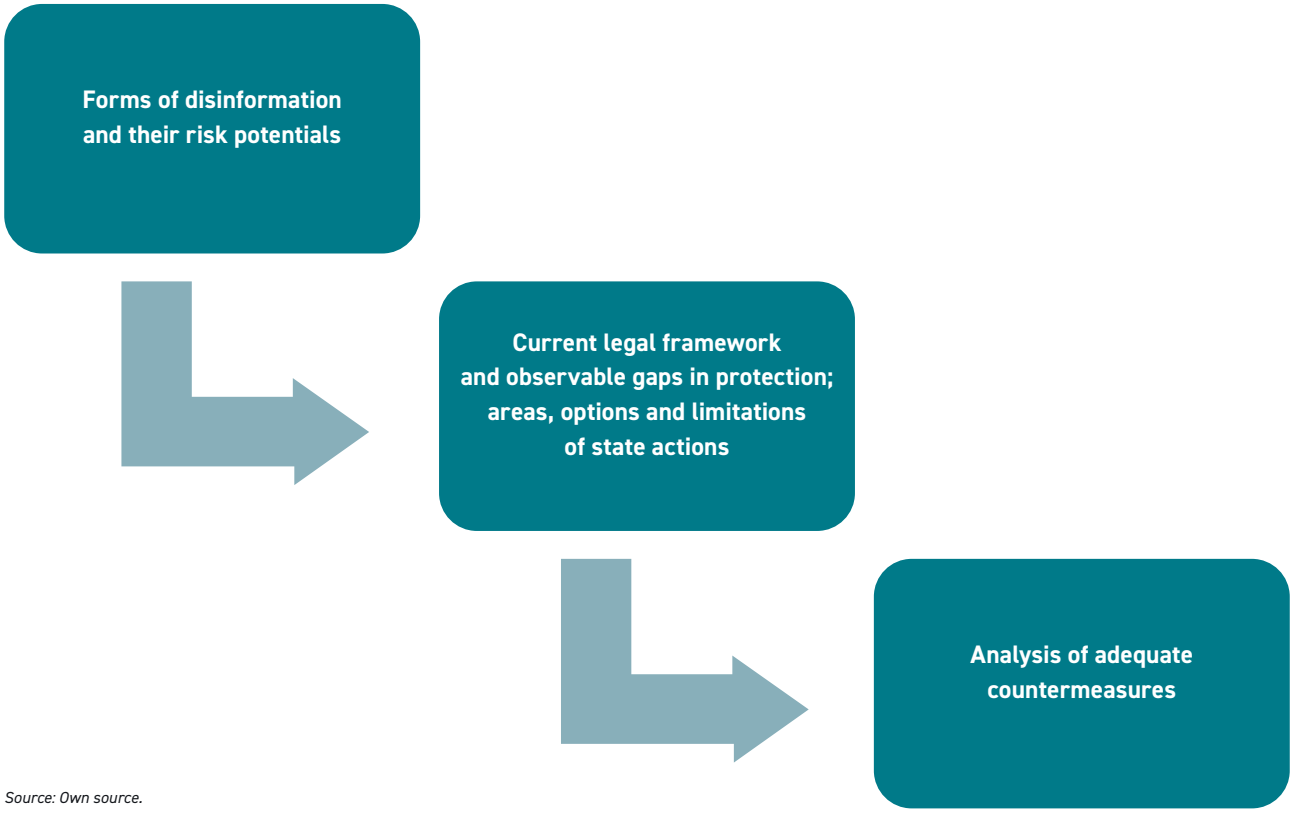
The second step (Chapter 3) examines the current legal framework for existing legal provisions against the realization of disinformation-based risks as well as initiatives that have developed on the level of co- and self-regulation that might work as counter forces. Here, the study continues the work of Möller, Hameleers and Ferreau[5] and their GVK study by highlight-

1 *Humprecht*, 2019, 1973.
2 Böckenförde, 1991, 92 (pp. 112).
3 See for instance *Mafi-Gudarzi*, 2019, 65 (68).
4 See Chapter 3.1.4.
5 *Möller et al,* 2020.

ing existing risk-specific gaps in legal protection with regard to the risk potentials identified in Chapter 2. Where we can identify gaps in the legal framework, we show options and limitations of regulatory interventions.

The third step (Chapter 4) focuses on regulatory starting points and instruments that might help in closing the identified gaps in legal protection. Traditional approaches in media regulations are usually less feasible in disinformation-related governance in view of the basic principle in public communication that it cannot and must not be the government's role to determine whether something is true/not true or a desired/undesired opinion. In this case – if action is even indicated at all – approaches have to be used that are independent from State action and rather work procedural than content-related , or alternative forms of content- and technology-related governance will have to be developed. Besides measures that enable or support public discourse, we have to consider countermeasures that are able to support the integrity of information.

Fig. 1: Schematic overview on the course of the study



Source: Own source.

The study has been conducted between November 2020 and May 2021; literature and initiatives through May 20, 2021 were taken into account. The team of authors would like to thank the wider group of legal experts involved in the project (Martin Fertmann, Amélie Heldt, Matthias C. Kettemann) as well as the student researchers at the HBI (Max Gradulewski, Rike Heyer, Maximilian Piet, Jan Reschke, Leif Thorian Schmied, Grace Ubaruta, Anna Zapfe).

6 *Katsirea*, 2018, 159.

2. TYPES OF DISINFORMATION AND THEIR SPECIFIC RISK POTENTIALS

The first objective is to provide an overview of the types and the criteria of disinformation used in academic discussions, to be, on the one hand, able to distinguish the subject of study from other phenomena, but also to critically question the usability of these criteria from a legal perspective (2.1). Based on the partially observable, partially rather hypothetical risk potentials of disinformation, the affected legally protected rights and constitutional goals are identified and described (2.2) to subsequently develop a legal approach to relevant dimensions of disinformation (2.3).

2.1 DISINFORMATION: TERMINOLOGY AND APPROACHES TO TYPOLOGY

2.1.1. State of research: Types of disinformation and distinctions from other phenomena

The scientific discourse comprises a series of approaches aiming to define and categorize forms of "problematic information."[7] The *problematic* aspect is always that an information is not concise, misleading, incorrectly allocated, simply wrong or made up. The forms of problematic information are becoming more differentiated in scientific discourse after that, depending on the intention of the person who shares it, on his or her motif, on the approach used to affect recipients or on the type of technical resources and measures to increase the visibility or reach.

Intent to Deceive

To that end, the first important point is the question whether the person sharing false information *intentionally* aims to deceive or mislead. If it happens unintentionally, by accident or negligently, the term used is "misinformation".[8] This category includes technical errors in research, unclear scenarios or cases of doubt, imprecise or ambiguous statements. It also includes playful, funny or ironic statements with satiric intent or that aim to put a hoax out into the world. In the latter cases, individuals sharing the information know that it is untrue but use it as an instrument for societal criticism. This type of misinformation specifically does not aim to deceive. The sharing of conspiracy stories can be seen as structurally similar: In these cases, the individuals making the statements are usually convinced that the information they share is true.[9] With regard to their point of view they aim to spread the *truth* and try to compel others of it. This usually excludes any intent to deceive (regarding the issue of using intent as a criterion see Ch. 2.1.2).

However, if the person who shares it knows that the information is untrue and aims to mislead the recipients, the information is seen as classic disinformation. Disinformation is the *intended* sharing of information that is untrue.[10] Cases when the individuals know that the information is wrong and share it anyway but do not have any intent to deceive or if the content does not have deception potential, are different. Such a lack of intent to deceive may point towards a different underlying motive, such as the desire to attract attention, commercial gains, or the proliferation of such an information to specifically differentiate oneself from such content. However, even if there is no intent to deceive, the person making the statements can, as a rule, not completely foresee the effect of the information on the recipient's side. Hence, the person making the statements usually also accepts the possibility of being misleading. This is seen as sufficient for the notion of disinformation.[11] In cases where an intent to deceive exists while the statement itself does not have the potential to do so (e.g. because the recipient can recognize the falsehood without effort or cannot understand the statement at all), the notion of disinformation should not apply, since it also requires the possibility that the information actually will succeed in misleading.[12] False information that does not disinform is therefore not disinformation.[13] However, the aspect of the potential of a state-

7 *Jack*, 2017, 1.
8 *Zimmermann/Kohring*, 2018, 526 (535).
9 *Fallis*, 2015, 401 (411).
10 *Möller et al*, [fn. 5], 11.
11 See *Fallis*, Library Trends 2015, 401 (410)
12 See in detail *Skirbekk*, 2016.
13 In reference to the phenomenon that the individual who makes the statement sometimes does not even want to mislead, but primarily want to polarise society, see *Bayer et al*, 2019, 27.

ment to deceive – regardless of the person's subjective intent to deceive – has not completely established itself as a component of the definition of disinformation. This will have to be further addressed from the legal perspective (see Ch. 2.1.2).

In some cases, academia takes a closer look at the different motives of the person sharing false information: The sharing of disinformation may be based on ideological or political motives; in other cases, the intent is driven primarily by financial motives, for instance, if it is possible to monetize information that draws a special amount of attention. In other cases, the sharers have neither ideological nor economic motives; their sole intent is to cause trouble (trolling).

Lack of Truth

The linchpin of virtually all definitions of disinformation is the *lack of truth* of a statement; it is untruth which turns information into *dis*information. In this context, it is frequently overlooked that truth as a term certainly has different (philosophical) schools of thought[14]: While metaphysical approaches presume that there is only one objective truth that applies to everyone and everything, constructivist approaches establish a world in which truth is constructed by every individual or by collectives. In other words, there is no absolute truth; instead, a large number of subjective and intersubjective truths exists. In this case, reality is that part of the world where subjective truths are in alignment. To that end it may make sense socially that parts of society develop their own practices of reality construction, so that for instance the actual absurd decision of a parliament to accept human-created climate change (whether it happens or not is of course not within the control of the parliament), holds an important function that marks a common understanding of reality and develop future action from that common ground. The question of how truth is constructed in society also plays a role in research on conspiracy narratives.[15] However, this strand of research itself struggles with the construction of the binarity true/untrue and must reflect that it creates its own object - "conspiracy" - through the questions of its empirical studies.

In scientific interpretations of disinformation, this truth/untruth complexity is primarily reflected in attempts that use the different levels of deviation of a statement from reality, if these deviations are provable. In fact the respective levels attempt to identify different degrees of falsification[16]: The deviation of a statement from an objectively observable state appears to be particularly easy to falsify (whereby this result is constructed through discourse). Completely made-up statements and attributions or alleged contemporary documents created by manipulating photographic, audiovisual or sound recordings are also part of this category ("manipulated content").[17] Depictions that show only excerpts of actually observable matters and can thus generate a certain perception of reality on the recipient's side while leaving out other important information or context, on the other hand, appear to be less "false." *Wardle*, but also *Egelhofer & Lecheler* describe, among other things, categories of such partially untrue "information disorders":[18]

"False connection": Using lurid or emotionalizing headlines, images or captions to attract attention, while the entire piece has nothing to do with the headline or the images or at least not directly. Common examples are clickbait formats.

"Misleading content": Content is moved into a certain interpretative direction through the misleading choice of photos or selection of image details. Given that the cognitive processing of images occurs more rapidly than the one of text, an image that tends to state something can affect the understanding of the overall contribution.

"False context": By leaving out the original context or by adding information, principally truthful information is changed in terms of its informative content. Examples are falsely referencing a source or the date an image was taken or quotes that have been taken out of their contexts.

"Imposter content": In these cases, the credibility of information is reinforced by adding known names, photos or brands.

Such types of "partial truths" are not out of discussion with regard to the existing definitions of terms. In respective examples, current approaches go back to the aspect of intention: The selective removal of the "whole truth" occurs intentionally to cause a certain intended effect with the recipient (i.e. perception of truth).

14 *Stapf*, 2021, pp. 105 (incl. further ref.).
15 Here, reference is often made to Bourdieu's distinction between doxa, orthodoxy and heterodoxy, see e.g., *Schink*, 2020.
16 *Tandoc et al*, 2018, 137 (147); *Egelhofer/Lecheler*, 2019, 97 (98).
17 E.g. deep fakes, that depict events that have not really happened by using technical processes, are new forms of media manipulation. In addition, a series of analogue and digital tools exist to modify and combine media content for the purpose of deception.
18 *Wardle*, 2018, 951 (953); *Egelhofer/Lecheler*, [fn. 15], 97 (98); for further types, see *Kapantai et al*, 2021, 1301.

However, the suggested truth that can be derived from such a statement is extremely variable – on the one hand from the individual recipient's point of view and from the perspective of a (functionally objective) observer on the other hand. Moreover, this depends on a number of linguistic, semantic and logical aspects, e.g. selecting merely specific information and leaving out relevant context information, accompanying circumstances, important prior and subsequent facts, but also co-mingling factual information with own assessments and subjective evaluations. Such actions can turn a statement, from the observer's perspective, into a mix of objectively describable facts and subjective perceptions and assessments with the opposite party (false implicatures).[19] Especially statements that contain externally verifiable core content that goes hand in hand with forms of personal insult, revelations, criticism or pranks comprise hybrid information that cannot always be completely decoded and respectively critically reflected by the recipient. On the other hand, if expressions of assessments and opinions are clear and do not consist of factual core, it must be presumed that from the viewpoint of communication science such statements are (no longer) disinformation.

Special Claims to Truth

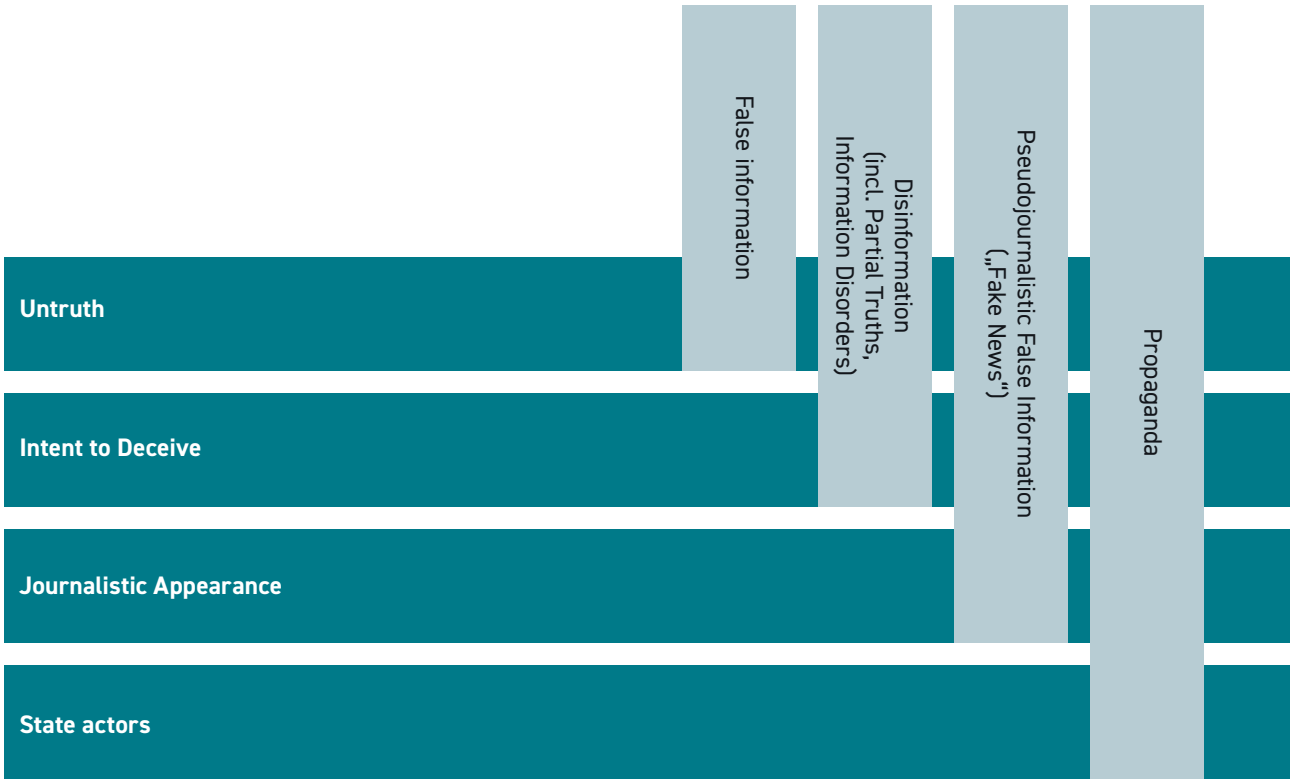
In addition, the style and appearance of information is often allocated to truthfulness: Journalistic statements and statements that appear to be journalistic seem more trustworthy and more credible to many recipients, because based on such signals they presume that the information specifically claims to be truthful.[20] This (supposed) truthfulness results in trust, amplifying the effects of deceit. This does not mean that statements without such an objective claim to truth are less effective. In particular, statements of close friends and acquaintances are principally given more relevance by recipients.[21] Information that appears as statements with special claims to truth has been discussed under the keyword "fake news" in recent years.[22]

The shared opinion is that "fake news" is a form of disinformation, whereby the term is simultaneously used as a description of a manifestation of disinformation and as a label.[23] Fake news as a genre of disinformation comprises the intentional generation and distribution of disinformation designed as a journalistic statement. As an ideologically charged label the term fake news is primarily used to discredit or delegitimize classic news media that allegedly report untruthfully or biased. According to opinions in scientific discourse, the term fake news should only be used to describe this strategic use, if it is used at all.[24] Given this background, the argument is that terms which should better be used in the discussion are pseudo-journalistic false information or current disinformation.[25] The central aspect of pseudo-journalistic forms of disinformation is that it has an apparently established medium as a source that is based on journalistic principles. Hence these terms do not apply to statements the source of which is clearly not a (pseudo) journalistic outlet, such as statements made by private individuals, even if the information is related to current events. Alleged witness statements are marginal cases, given that the presumed observers in these cases become live reporters who may be given greater credibility – in particular if these eyewitnesses corroborate what they have (allegedly) experienced with sound or video recordings.

Another special sub-type of disinformation is propaganda, which is characterized by the fact that in these cases, especially governmental or political players pursue a specific agenda through systematic disinformation. The aim is once again to deceive or mislead the recipients. This either happens through the biased selection of truthful information distributed by official government agencies ("white propaganda"), or through the intermingling of truthful and dubious statements ("gray propaganda"). Propaganda also includes forms of distribution of misleading information for which the source of the statement is falsified or concealed ("black propaganda").[26] An additional special aspect of this type of disinformation is that it appears to come from the scope of governmental or political centers of power and that it is systematically used for the purpose of significantly manipulating larger groups of populations in their points of view and attitudes (see Figure 2).[27]

19 *Fallis*, 2014, 138.
20 *Holzer/Sengl*, 2020, pp. 161.
21 *Samuel-Azran/Hayat*, 2019, 71.
22 Cf. for instance Tandoc et al, [fn. 15], 137; *Zimmermann/Kohring*, [fn. 8], 526; *Roozenbeek/Linden*, 2019, 570; *Egelhofer/Lecheler*, [fn. 15], 97; *Chambers*, 2021, 147; *Bakir/McStay*, 2018, 154; *Gelfert*, 2018, 84.
23 *Zimmermann/Kohring*, M&K 2018, 526 (527)
24 *S. Wardle/Derakhshan*, 2017, 16; *Vosoughi et al.*, 2018, 1146 (1).
25 *Zimmermann/Kohring*, M&K 2018, 526 (527)
26 *Jack*, 2017, 7.
27 See *Möller et al*, [fn. 5], 30. A blatant deviation from factuality as described there, however, is not a mandatory criterion of propaganda; cf. *Tandoc et al*, [fn. 15], 137 (146).

Fig. 2: Types and Criteria of "Problematic Information"



Source: Own Source.

Special Forms of Distribution

A special aspect of digital disinformation – which also applies to digital communication overall – is the option to generate, publish and distribute information easily and on a massive scale, for instance via email, messenger services or social media such as social networks, video sharing platforms or micro blogging services. The (re-)distribution of a statement, once it has been published, cannot only be handled by the actor who makes the original statement, but also by readers and recipients of the initial statement. By forwarding, liking, sharing and linking it, it is possible to attain vast reaches in a short period of time, while the content of the statement usually does not change; it "runs along" as the target reference of the proliferated information. The redistributing users often do not have any intent to deceive but may already have been misled by the original disinformation. If an information triggers major attention and a large degree of redistributive interactions, this circumstance is frequently fed into platforms' selection and prioritization algorithms as a signal. Based on this, the visibility of the respective content is prioritized and shown or recommended more frequently in personalised media environments, which in turn increases the reach of and the interaction with the content.[28] As a result, feedback loops are evolving, in which initially "successful" content and the selection logic of the respective platform reinforce each other. As a result, sometimes viral distribution reaches and speeds are achieved, which intensifies the respective scale and impact effects on public discourse (see Chapter 2.2.4).

Besides organic forms of information distribution and its interaction with selection and recommendation algorithms on user generated content platforms, the person making the statement or a third party may also help the statement attain artificially increased reach. To do this, networks of accounts that have been created specifically for this purpose are used that quickly like, share, forward or comment on certain content. These accounts ("sock puppets")[29] can be manually operated by the respective account holder or they can work semi- or fully automated. Thanks to these types of (fake) account networks, it is possible to fake the degree of user interest and cause an artificial amplification of the content by the platforms.[30] Another option to evoke attraction regarding a particular content is the targeted interaction with special groups

28 *Rader/Gray*, 2015, 173.
29 *Lewandowsky et al*, 2017, 353 (2).
30 *Puschmann*, 2020, pp 544.

among the population or specific groups of interest. Many platforms with user generated content allow paying customers to publish ads on the respective platforms and to select as well as limit the group of addressees for the information in the ad in a more or less concrete manner, e.g. based on age, gender, place of residence, preferences or political views (micro-targeting).[31] With regard to disinformation, this approach enables targeted communication to those individuals where the highest acceptance rates are expected, probably resulting in better sharing and proliferation.

However, when it comes to disinformation, such distribution options are not limited to one platform. In fact, systematically planned information campaigns can be observed that launch cross-platform campaigns (often customized for the respective platform) and that pursue specific agendas with identical or similar information. Another special case can be seen in (pseudo-)journalistic sources, such portals with news and typical formats of news reporting, which publish disinformation and subsequently generate reach through links to and adoption by social media platforms. Linking and sharing such articles on social media platforms result in accessible short posts with title, image, summary or teaser and a link to the external source that can easily be liked, shared and forwarded on these platforms.

An indirect form of redistribution can be seen in the adoption of information or its distribution on social networks through established media outlets. For instance, a much larger number of people have learned of Trump's tweets through media reporting than the 45th President of the United States had followers on Twitter. Traditional media also observe the developments and topics on social media because it is very easy to access discussions among parts of society there. Especially when it comes to information that is distributed quickly and vastly, this circumstance already shows some news value and a (presumed) reporting interest for this circumstance may exist. From the perspective of reporting it does not matter whether the virally distributed or emotionalising statement is true or false in the first place. The circumstance alone that a certain piece of information is shared a lot may have a value in journalistic reporting. As a result of the adoption of an (untrue) statement in news reporting the reach of the respective information is once again expanded. Hence, also traditional media play a relevant role in the redistribution of disinformation.

2.1.2. Central Criteria of Current Categorization Approaches as Legal Starting Points?

The overview on understandings and types of disinformation indicates that two central criteria appear in virtually all definitions: A statement is partially or entirely not true and this statement is made intentionally to deceive or mislead third parties. In subcategories one finds further criteria, such as a certain appearance (e.g. fake news), a specific player (e.g. propaganda) or a certain motive (e.g. ideological or financial). From the perspective of communications science, this approach defines a communications phenomenon and can be used accordingly to describe the object of investigation. From a legal or regulatory perspective, these starting points prove to be problematic with regard to their operationalization in implementation and from the perspective of legally protected rights. These definitional starting points seem to be of limited help when adopted in legal or legislative contexts.

Combination of Intent to Deceive and Untruth as a Legal Starting Point?

A cross-sectional issue arises from the definition-based assumption that the intent to deceive and the untruth of an information always appear together: Someone wants to mislead third parties with factually wrong information. However, this approach overlooks large parts of distributing false information (a) by persons who have already accepted the perceived information as a fact and now distribute it as allegedly true information to their contacts and (b) by persons who are aware of the fact that the information is untrue, but share content out of amusement or as irony and (c) by persons who adopt untrue statements and share them out of financial gain and (d) by persons who have recognized or know that the information is untrue, but use the shared information primarily as a way to state their opinion, e.g., to show their world view or political attitude.[32]

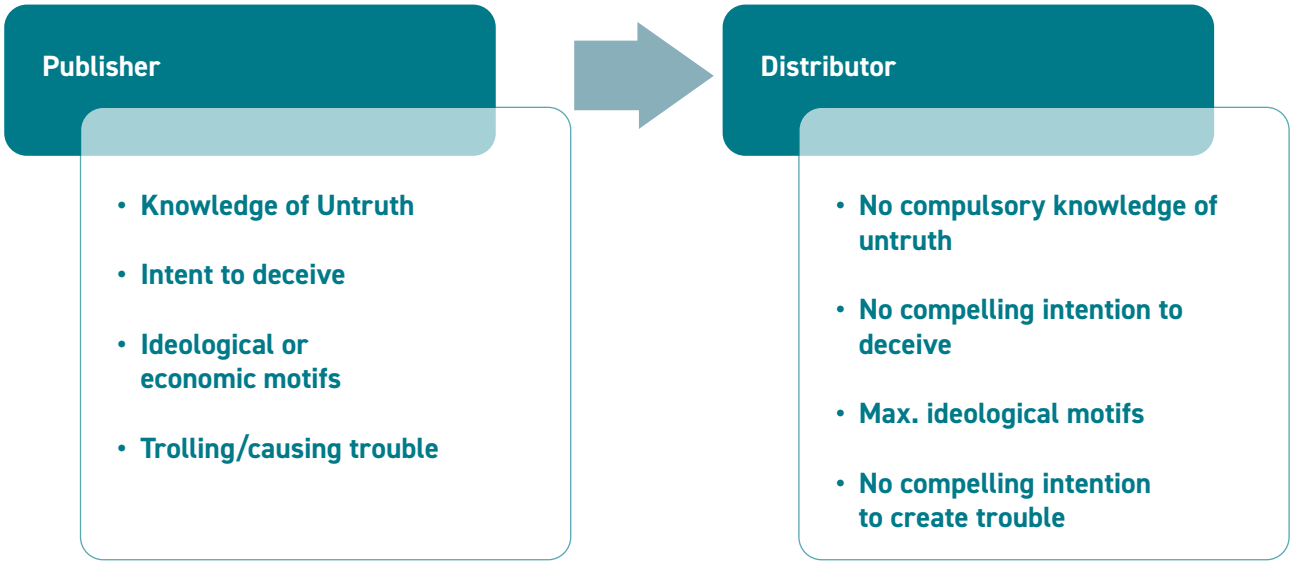
> „This indicates that people in digital communication environments are far from hapless victims of manipulation but find ways to socially verify information, if so inclined. More fundamentally, instead of predominantly informational functions, the use and distribution of disinformation follows other functions, such as the signaling of social belonging (...), political

31 See *Barbu*, 2014, 44 (45).
32 *Puschmann*, [fn. 29], 540; *Duffy/Ling*, 2020, 72; *Allcott/Gentzkow*, 2017, 211.

partisanship (...), or an impulse to challenge commonly held beliefs or social values (...). Disinformation is thus not a driver of social or political divisions. Instead, it is an expression of them."[33]

In all four cases, an intent to deceive does regularly not exist. Nevertheless, these groups support the distribution of false information and thus increase the potential of its effects. As a result of the change in players from producers to distributors and redistributors, their intentions and motives may change significantly, however, without reducing risks for legally protected rights on the recipient's end.

Fig. 3: Changes of Intentions and Motives in the Chain of Actors



Source: Own Source.

Intention to Deceive as a Legal Starting Point?

Regardless of the problem of coupling the two explained characteristics, the proof of an intention is not trivial as far as law is concerned. An intent to deceive as the prerequisite for the affirmation of disinformation would be no exception. On the one hand, to presume an intention to deceive, evidence would be required that the person making the statement is aware of the untruth of the information he/she is sharing. Especially in the described categories of half-truths and information disorders this is not easily possible, because the information as such is not (completely) untrue, and potentially misleading effects of distorted or decontextualized information do not materialize until reaching the recipients' side (false implicaturs). On the other hand, the intention would have to include the possibility of *being able* to deceive the recipients of the information. This refers to the awareness of the potential to deceive of the self-distributed statement. If the person sharing it doubts the possibility that the false information could compel others of its truth and that the statement is highly likely to be recognized as disinformation, it can no longer be presumed that a deception will actually occur.

If one were to assess intently acts from the perspective of a neutral third party or the viewpoint of an objective recipient, as is the case in law, this approach cannot clarify all cases of potential intent to deceive. For types of manipulated content in which individual contributions or entire content services are produced with so much effort that they look particularly credible, one might suggest – from an objective point of view – that the provider indeed has an intent to deceive. But with regard to the first person who redistributes the information, an objective perspective as proof of an intent to mislead would already meet its limitations. Moreover, in the case of cascades of individuals making statements by sharing information, this assessment would have to be applied repeatedly for each individual and, depending on the presumptive status of individual knowledge would have to either approve or prevent the distribution of identical content. Hence, intention-based criteria do not offer a tool against the *further distribution* of disinformation. It has also become evident that the potential

33 *Jungherr/Schroeder*, 2021, 1 (3) (incl. further ref.).

to deceive is inherent in false information, regardless of the existence of any intent to deceive on side of the expressing person. Especially the risk for legally protected rights of the recipients, though, is a central starting point for regulatory interventions from the State's point of view (see Chapter 4).

Given the multilayered content of statements, the determination of intent also has to take into account external circumstances. In the complete absence of context, the proof of an intention behind statements is usually not more than an allegation. Alternatively, including the context of an information, such as the frequency, other statements made by a person or other information about the political or ideological attitude, would lead to the result that the aim is no longer the actual relevant statement at the heart of the matter. In this case law would become rather attitude-based criminal law which takes into account the general world view of a person to make a decision on legal sanctions. All in all, an intent to deceive can only be proven in exceptional clear cases as a functional criterion for disinformation – and if so, primarily as an approach to determine the degree of guilt.

Untruth as a Legal Starting Point?

The second central criterion in the current definitions of disinformation is the circumstance that the stated expression is (partially) untrue. As described above, truth is primarily a construct resulting from interpretation, and reality as commonly shared truth is primarily the result of social (communicative) negotiation processes.[34] Any form of statutory, government or regulatory provision or definition of truth undermines these negotiation processes and defers the power over the definition of truth from the area of societal self-understanding in the direction of a truth imposed by the state (see Chapter 3.3). Even the fact-finding process by courts in individual cases does not represent government-defined truth but reflects the duty of the court to search for the truth. Hence courts are part of and observers of the societal negotiation process of truth, which they then reconstruct in an orderly, independent fact-finding process. The result of this process then creates the basis for the court's decisions in the context of its application and verification of law.[35]

From a legal perspective, another issue renders the scientific definition of disinformation even less useful: The freedom of expression guaranteed by human and constitutional rights frameworks encompasses all types of opinions and statements of facts. Opinions, in this context, are subjective assessments that are neither correct nor false, neither true nor untrue, but part of the expression of any individual's world views. There are ethically and morally questionable, absurd or even inhumane opinions – they all remain permissible statements with very few exceptions that are criminally impermissible. In expression-related law, the starting question for a legality check is therefore whether the statement in question is accessible to proof (statement of fact) or characterized by an element of opinion (statement of opinion). How a statement has to be interpreted, whether it is a statement of fact or the expression of an opinion, who is required to prove the truth of a statement and which requirements the evidence must meet, is subject to complex rules and heuristics. A fair – and constitutionally compliant – proceeding hence requires an elaborate assessment of the individual case taking into account the narrower context of the statement. The simple interpretation of a relevant (untrue) statement as objectively false information will regularly reach its limitations when assessments and claims of facts are so intermingled that they can no longer stand on their own – which happens regularly:

> "Even claims of facts – as constructs of reality – inevitably represent the result of a selection, interpretation and reflexive process that cannot be bypassed (even if this isn't necessarily a matter of awareness under every aspect!) by a concrete (expression) subject in its individual references to reality."[36]

In addition, disinformation, as shown above, frequently uses specific forms to distort true facts, which can also result in intended understandings at the recipients' end (false implicatures), for instance as a result of the selective choice or the selective omission of facts, without using an objectively untrue claim.

Regardless of this, the criterion of untruth may cause legal result problems in special communicative constellations: It is common view in jurisprudence that in certain areas of life, discussion is particularly emotional, complex facts are sometimes abridged or presented in a polemic manner or exaggerations and glorifications are on the agenda regularly, for instance in political debates – in particular in times of election campaigns – as well as in commercial communications and

34 *Habermas*, 1973, pp. 211.
35 *Jung*, 2009, 1129 (1130).
36 *Jestaedt*, 2011, marginal note. 37. [Translation by the authors].

advertising. However, even exaggerations and polemic statements are regularly permissible statements. Even publicly expressed untruths that later prove to be "white lies," for instance because of a news embargo or the application of pressure, may be untrue without showing any specific degree of wrongdoing. The definitory starting point that public statements are always objectively falsifiable and thus can be separated into true/untrue, is thus a rather artificial point of view that might be used in fruitful ways for theoretical or methodical approaches in communication sciences. However, where legal measures (and constitutional restrictions) such as bans, deletions or sanctions are linked to the presence of disinformation, the central criteria of current scientific definitions are not very valuable for legal and regulatory approaches to disinformation.

2.1.3. Derivation of Legally Relevant Dimensions of Disinformation

The above-described definitions of disinformation refer to common criteria that are only somewhat suitable for the appearance-related differentiation and distinction from other, similar phenomena as well as starting points for regulatory approaches. This is no cause for concern and can be attributed to the disciplinary, system(logic)-immanent perspective on a phenomenon[37] – as long as a control-oriented sciences such as law do not adopt such definitions unquestioned.

Any form of state governance of (dis)information[38] necessarily touches the scope of basic communications rights as guaranteed by the European Convention on Human Rights (Art. 10 ECHR), the Charter of Fundamental Rights of the European Union (Art. 11 CFR) and the German Constitution (Art. 5 Sect. 1, 2 GG). Impairments or interventions in the scope of application of these rights therefore require justification. Here, a special feature applies to Germany's constitutional right to freedom of expression: The German Constitution does not consider statements protected by Art. 5 GG that are knowingly untrue.[39] Certainly, one of the reasons for this is that it would appear to be intolerable to many if the Constitution protected the denial of the Holocaust. Another is that of the three functions for which communication typically appears to be particularly worthy of protection – personality development, formation of democratic will and truth-seeking – the latter plays a comparatively minor role in the Philosophy of Law. However, this does not mean that for instance the prohibition of lies would be compliant with the constitution regardless of concrete dangers to legally protected rights.[40] Art. 10 ECHR protects the expression of false claims, but allows for more intense restrictions.[41]

The central legal starting point for statutory interventions in (dis)information must therefore be the dangers of problematic statements for legally protected rights (see Chapter 2.2).[42] This change of perspective puts the aspect of risk to legal interests and thus the potential effects on the recipient's side in the focus of any legal consideration. Moreover, this changes the focus of the current definitions away from any intention to deceive on side of the expressing person to the potential effects on the part of the receiving individuals, where the recipients are usually the persons affected by the detrimental effects. However, there are also constellations in which those affected in their legally protected rights do not belong to the group of recipients (see Chapter 2.2). Recognizing the threats to legally protected interests enables regulatory approaches that tie in with possible or probable consequences of communication without the need to define the specific form or type of the relevant communication in more detail or using criteria that are difficult to operationalise. On the other hand, interventions due to (too) abstract risks are particularly difficult to justify. In this respect, it is important to understand the effects of (dis)information on legally protected rights. By now, it is already graspable that this necessary approach will increase complexity in view of a variety of possibly affected rights and interests as well as the context-dependent variance of types of disinformation.

By focusing on the risks, less central criteria of the definitions described above come to the fore in legal considerations: In particular, the concept of the potential for deception, the associated responsibility for statements with a special claim to

37 Communication sciences need a definition as a point of departure for the description of the phenomenon to be researched, in distinction from other forms of communication, while law describes phenomena on an abstract level using general criteria that must be applicable to individual cases to be able to tie legal consequences to it.

38 A consequence of the deviating legal understanding of disinformation is that one must no longer speak of disinformation in a legal study or one needs an alternative definition. Here, we use the working definition given at the end of this section.

39 BVerfGE 54, 208 (219); 85, 1 (15); 61, 1 (8); 90, 241 (247).

40 *Saliger*, 2002, pp. 102. Such a measure would also have to be checked against the freedom of action in Art. 2 Sect. 1 GG. In addition, freedom of information is focusing on the source of information to which the access is impaired, and not on the notion of opinion within the context of the freedom of opinion.

41 European Court of Human Rights, 6.9.2005 – 65518/01, Slg 05-VIII Rn. 113 – Salov/Ukraine.

42 *Buchheim*, 2020, 159 (pp. 166).

truth and the measurement of relevance due to the supposed large number of supporters of a certain statement can be seen as transferable starting points. It is neither the malicious intent to deceive nor the untruth that are prerequisites for governance measures, but specific aspects of information, its appearance and dissemination in case they show risks for legally protected positions.

A particular challenge here is the determination of a legally relevant potential for deception, since empirical research has so far only been able to produce limited insights as regards the scope and effects of disinformation.[43] At most, findings regarding short-term effects are to be found[44]; whether and to what extent false information has an effect on attitudes and impulses for action in the medium and long term is still underexposed. The reason for this can also be found in methodological challenges of identifying (mono-)causal connections between a certain information and a specific attitude on the receiving side in complex media environments and even more complex opinion-forming processes (see Chapter 2.2.2). An objectification of any recipient horizon is only possible under difficult conditions, since the information literacy of individuals shows high variances and the individual information processing is (also) dependent on the respective platform and information context.

Here, too, the challenge appears that expressions of opinion are particularly protected and, as a rule, should not become subject of legal measures. Exceptions to this rule are criminally relevant statements, in particular with regard to the protection of honour or concrete violations of personality rights (see Chapter 3.1.1). Within these limits of what can be said, dissenting and discriminatory opinions that lie outside the socially appropriate, such as xenophobic, sexist or racist expressions of opinion, must also be tolerated. The European Court of Human Rights (ECtHR) expressly argues with the harmful effect of such statements when it states that Article 10 ECHR also protects statements that "offend, shock or disturb".[45] If permissible expressions of opinion are mixed with factual assertions, difficult-to-resolve mixtures can arise, which must be interpreted and differentiated in their respective contexts in individual cases. In view of the German Constitutional Court (BVerfG) the freedom of expression has to prevail in such cases of doubt:

> "The differentiation between opinions and factual assertions can of course be difficult, because both are often combined with each other and only together make up the meaning of an expression. In this case, a separation of the factual and the judgmental components is only permissible if the meaning of the statement is not distorted. Where this is not possible, the statement must be regarded as an expression of opinion in the interest of effective protection of fundamental rights and must be included in the protection of freedom of expression, because otherwise there would be a risk of a substantial shortening of the protection of fundamental rights [...]."[46]

Thus, statements about (alleged) events without value judgments can also contain statements about a person's characteristics (e.g. a prominent person allegedly beats his/her children; a political personality allegedly shows a Nazi salute). The challenge here lies in determining the descriptive meaning of alleged (and unverifiable) events or attributing alleged but decontextualized actions or statements.

When determining the potential for deception, it is also important from the point of the recipient's view to take into account the degree of credibility shown to a statement. For the more truthful recipients deem an information, the more uncritically they will use and process it.[47] Decisive for the credibility of an information are the aspects of its textual and visual appearance, the attribution of trust credited to the source of the information and – where the source and distributor of the information fall apart – the sender of a forwarded information.[48]

A person making a statement can assert a special claim to truth, for example, by claiming or supposedly testifying contemporary witness statements or by replicating statements of persons that are obliged to the report of true facts. The latter example refers - above all - to journalistic reporting: If the impression arises on the recipient's side that a statement is

43 Furthermore, where empirical statements are made, a binary concept of truth was implicitly assumed. A statement was either true or false; from a legal perspective, there are often hybrid forms, so that the empirical results are not precisely 100% transferable to the legal perspective (see Chapter 2.1.2).

44 *Landesanstalt für Medien NRW*, 2020, pp. 25.; see *Thorson*, 2016, 460.

45 ECHR, 23.09.1994 – 15890/89 marginal note. 30 (a) – Jersild/Denmark; see also ECHR, 07/12/1976 - 5493/72 marginal note 49 – Handyside/United Kingdom.

46 BVerfGE 90, 241 (248) with reference to BVerfGE 61, 1 (9); 85, 1 (pp. 15). [Translation by authors.]

47 *Högden et al.*, 2020, pp. 78.

48 *Landesanstalt für Medien NRW*, [fn. 43], 28; *Högden et al.*, [fn. 46], pp. 78.

made by a person or an institution that is obliged to apply duties of care in reporting as truthfully as possible and that has to clearly identify expressions of opinion or value judgments, then this has an impact on the potential for deception. Where it can be assumed - from an objective point of view - that the person making a statement deliberately uses a (pseudo-)journalistic appearance, he or she must also be measured against the duties that apply to truthful journalistic communication (see chapter 3.3). Incidentally, this approach does not exclude the possibility that a potential for deception may arise even in the case of untrue information *without* journalistic appearance.

A final relevant aspect of the potential for deception of information is its dissemination: if a statement is not received, its risk potential cannot materialise. With increasing distribution, high reach and many individual contacts, the visibility and reception of the expression increases. A basic potential for deception is thus theoretically multiplied. From the point of view of the person making the statement, its dissemination cannot be fully anticipated; in some cases, the amplification of information is not necessarily intended, but is rather the result of the selection logics of platforms (see above). In this respect, a legal assessment would depend on an retrospective analysis. However, if a person deliberately and significantly promotes the dissemination of a dubious statement, for example through the systematic use of automated accounts, this could be used as a legally relevant criteria (see chapter 3.3 below).

Against this background, we use the following working definition for disinformation for this study, which differs significantly from the current non-legal definitions:

Disinformation describes utterances,

(1) the truth of which can be doubted with good reason,

(2) which can easily be disseminated and shared,

(3) which due to the person making the statement or due to their design claim to be truthful from an objective recipient's perspective, and

(4) which impair legally protected rights.

2.2. SYSTEMATIZATION OF RISK POTENTIALS OF DISINFORMATION

As shown, the necessary starting point of a legal perspective on disinformation must focus on its potential to infringe legally protected rights and thus the potential effects on the recipient; it's this side where the relevant effects of false information occur and thus may have adverse effects on protected legal interests, positions or legal principles.

2.2.1. Threats to Legally Protected Rights as a Starting Point

The central point of departure of this analysis is the disinformation-specific risk potential for constitutionally protected rights and freedoms. The purpose of the following is to systematically identify such risks and analyze them for their suitability as legal criteria. After all, when governance measures aiming to combat disinformation regularly represent an infringement of the freedom of expression of the person making a statement, they appear justified only if a risk of not meeting legitimate constitutional or human rights objectives and if the legal intervention helps to promote the realization of these goals[49] (see Chapter 3.4). Herein, we differentiate between risk potentials for individual rights, for supra-individual rights and for societal legal interests and principles. Within these categories, one can also differentiate between direct and indirect risks.

49 *Hoffmann-Riem*, 2002, 175 (184).

From the perspective of communications science, this is a special point of view, given that protective goals set by law have to be identified and then effects have to be assessed with a specific view on risk potentials for these goals. With regard to the effects of disinformation[50] only few meaningful empirical studies are available to date[51], and they are mostly based on the above-described communication science-oriented definition of disinformation. Some of these contributions summon risks to democracy caused by disinformation, for which no further evidence is provided[52], but that rather use it as an argumentative legitimization for subsequently proposed counter measures.[53]

> "For a topic like disinformation that elicits such strong fears, the empirical basis of the debate is rather thin. [...] The few studies that have empirically tested the reach of disinformation consistently find this reach to be severely limited [...]."[54]

Thus, the following chapters can be seen as both, a compatible basis of legally informed research on disinformation effects in the future, and a starting point for a debate within legal sciences on what (societal) risks actually require an intervention.

2.2.2. Individual-related Risk Potential

Individual rights and freedoms may be affected if disinformation affects the process of the individual opinion formation in a way that it impairs the individual autonomy or the freedom of forming a political will. In addition, disinformation has the capability of endangering knowledge building as the basis for the formation of a free will. If wrong information is related to certain persons or groups, this may also be seen as libel, showing specific intersections with so called "hate speech." Ultimately, such utterances may also result in danger for physical harm and threats to life, where disinformation is the source for indirectly damaging activities.

Infringement on Autonomy

Autonomy refers to an individual's ability to make independent decisions.[55] As an individual freedom it derives from the general freedom of action[56] and is a central element of the general right to personality, Art. 2 Sect. 1 GG in combination with Art. 1 Sect. 1 GG.[57] The general right to personality protects autonomy as a basis of self-determined development and maintenance of one's own personality.[58] A European equivalent for the general right to personality can be found in Art. 8 ECHR and Art. 7 CFR.[59] It is presumed that disinformation because of its potentially misleading effects sometimes can influence decision-making processes including the outcome of decisions, causing a risk of impairment of an individual's autonomy.[60] One unique risk has been identified in connection with specific online manipulations[61], such as those that arise from forms of micro targeting.[62] It is said that this kind of manipulation undermines the individual autonomy, since its misleading effects can cause persons to do things for reasons and purposes that are not authentically their own.[63]

However, for the presumption of an impairment on individual autonomy in a legal sense, narrow conditions must be met. Basically, the scope of protection of the general right to one's personality (as is the one in Art. 8 ECHR)[64] is open for new developments and can be expanded by case law based on the risks created by new societal and technological phenomena for the development of one's personality.[65] However, individuals as social beings living in communities are unavoidably

50 Overview in *Högden et al*, [fn. 46].
51 See for instance *Grinberg et al*, 2019, 374; *Guess et a*l, 2019, 1; *Barua et al*, 2020, 1; *Hameleers et al*, 2020, 281; *Zerback et al*, 2021, 1080; *Bail et al*, 2020, 243.
52 See, also, but further differentiating *McKay/Tenove*, 2020, 1.
53 *Zimmermann/Kohring*, [fn. 8], 526; *Turcilo/Obrenovic*, 2020, pp. 19.
54 *Jungherr/Schroeder*, 2021, 1-13.
55 *Susser et a*l, 2019, 8.
56 See *Di Fabio*, 2020, m.a. 130.
57 See e.g. BVerfGE 141, 186 (201 f.); the Court states that it is a "function of the general right to personality [...] to safeguard fundamental conditions that ensure that individuals can develop and protect their individuality in a self-determining manner [...]".
58 See BVerfG, NJW 2016, 1939 (1939 f.) (incl. further ref.).
59 Regarding the right to one's own personal development and autonomy cf. ECtHR, 23.03.2017 - 53251/13 m.a. 76 - A.-M.V./Finland; 19.05.1976 - 6959/75, P. 115 - Brüggemann and Scheuten/Germany.
60 *Craufurd Smith*, 2019, 52 (64) (incl. further ref.).
61 *Susser et al*, [fn. 55], (6). The authors define online manipulation as the use of information technology to influence the decision-making process in a concealed manner through targeting their decision-making weaknesses.
62 *Susser et al*, ibid., (6). However, the authors refer to online manipulation in general and not to disinformation-based manipulation.
63 *Susser et al*, ibid., (9).
64 *Marsch*, 2018, pp. 217.
65 Cf. e.g. BVerfG, NJW 1980, 2070 (2070).

constantly exposed to external influences, against which legal rights usually do not exist. Accordingly, the German Constitutional Court (BVerfG) states that the right to personality does not protect against "everything that could impede the self-determined development of personality in any way," since "no human being [can] develop his or her individuality independent of external circumstances and relationships [...]."[66] Hence, a certain pertinence must exist, which is not excluded in the event of exposure to disinformation but requires empirical evidence with regard to the actual effects. However, to date, only few studies of such effects in conjunction with disinformation have been conducted, so that it cannot be presumed with certainty that the distribution of disinformation leads to such an influence over the process of the individual opinion formation[67] that it would have to be presumed that it places the autonomy at risk.

This lack of empirical evidence also has to do with the fact that research would have to examine complex processes of opinion formations on the individual level, which cannot be readily reconstructed and which can have effects on entirely different levels (See Tab. 1).

Tab. 1: Types of Media Effects on Users

Type of Effect	Short-term Effects	Long-term Effects
Dissemination of Knowledge	Factual knowledge regarding specific current events	Background knowledge regadring the political system and the emergence of long-term political problems
Agenda Setting	Perception of the relevance of specific political issues (e.g. scrappage bonus, killer games, Afghanistan deployment)	Perception of the relevance of common topics (e.g. ecology, peace, social security)
Framing	Recognition of specific aspects of current political discussions, specific perspectives on one topic	Encompassing perspectives on political topics
Imparting of opinion climate	Perception oft he distribution of opinions regarding specific discussions	Perception of typical (popular) attitudes towards political topics
Persuasion	Changes in attitudes on the level of specific topics or individual politicians	Long-term political attitudes and preferences for political parties, attitude towards politics in general
Stimuli for Action	Intent to act and actual election choice, participation in specific situations	General behavior with regard to participation

Source: Hasebrink et al, Macht als Wirkungspotenzial. Zur Bedeutung der Medienwirkungsforschung für die Bestimmung vorherrschender Meinungsmacht 2009, S. 5 [own translation].

To date, no indications have been found that disinformation could lead to *changes in actual opinion* (in the table this would be on the effect level of persuasion). What could be shown, though, is that already existing views of individuals can potentially be reinforced as a result of disinformation (confirmation bias[68]).[69]

66 BVerfGE 141, 186 (202).
67 *Roozenbeek/Linden*, 2019, 570.
68 See *Nickerson*, 1998, 175.
69 *Guess et al* and *Grinberg et al* took sharing of disinformation as proxies: *Guess et al*, [fn. 50], 1 (2); *Grinberg et al*, [fn. 50], 374 (5); *Hameleers et al*, [fn. 50], 281 (298); *Zollo*, 2019, 1.

Freedom of Political Will and Opinion Formation

The potential impairment of the individual formation to form one's opinion simultaneously comprises a potential risk for the (more specific) freedom of forming one's political will. A democratic will arises from an open and free process to form a political will of each individual, whereby the autonomous freedom of establishing a will on the individual level transfers to the political community and ultimately forms the population's sovereign will of the people.[70]

Consequently, in addition to individual legal guarantees, Art. 5 GG contains a component of positive obligation – also in conjunction with the principle of democracy principle arising from Art. 20 GG[71] – wresith regard to the freedom of the process of political will formation.[72] After all, democracy repents the legal-organizational implementation of the self-determination principle[73], for which the prior free and open opinion formation process is constituent.[74] While Article 10 ECHR does not provide a parallel positive legal obligation,[75] topics of public interest still enjoy special protection also under this article. Hence, the European Court of Human Rights stipulates that a specific justification is required to legally infringe with the freedom of expression that limits statements on matters of public and in particular political interest.[76] Consequently, the freedom of political will formation is considered particularly protected on both the European and the national level.

The process of opinion- and will-formation is characterized by the quality of the accessible information.[77] The reception of disinformation may let objectively false content become the basis of individual formation of opinion and may unwarrantedly influence potential political convictions and decisions. If, in a political context, disinformation influences the autonomous formation of will of the recipients, this results in an impairment of the freedom of political will formation.[78] The question is, as of which form of expression, as of which frequency or as of which degree of influence it can be presumed that a regulatory intervention appears to be justified; here, the empirical evidence is similarly thin as the one discussed above. However, the protected right of freedom of political will formation is significantly more specific than the general decision-making autonomy arising from Art. 2 Sect. 1 in combination with Art. 1 Sect. 1 GG. It thus appears possible that the requirements with regard to the significance of influence may be reduced for very specific manifestations of potential influence, in particular when it comes to the manipulation of definite political attitudes.

Another risk for the fundamentals of opinion and will formation arise from a wrong knowledge basis. The almost complete lack of access barriers provides users with a gigantic choice of information online, which can principally contribute to a more self-determined acquisition of information and to a better participation in public discourse.[79] However, the lack of access barriers can also lead to a threat to knowledge due to the increased and far-reaching distribution of disinformation: On the one hand, disinformation, regardless of whether it is believed, creates "relevant alternatives" that can make it difficult for recipients to gain shared knowledge if they cannot identify these alternatives.[80] The ability to distinguish between truth and fiction is getting harder for recipients.[81] Moreover, due to their similar appearance in social networks, shared content, regardless of its claim to truth, is assessed less critically when it comes to the credibility of the sources.[82] In the worst-case scenario, the knowledge of the individual might become segregated from the basis of common knowledge that is societally shared. However, a reliable, realistic image of the world is a crucial precondition for the effective assertion of individual interests.[83] In this context it should be taken into account that in Germany, unlike in other countries, information communicated via intermediaries only represents a portion of the information diet and that they are rarely the only sources used.[84] On the other hand, the established media is increasingly adopting or reporting wrong information distributed on the Internet[85] - despite critical distinction and disclosure of the falsehood. This way, false information might get stuck in the heads of the recipients.

70 *Böckenförde*, 1992, 455.
71 BVerfG, NJW 1966, 1603 (1604).
72 *Löber/Roßnagel*, 2020, pp. 150, ref. BVerfGE 82, 272 (281).
73 *Böckenförde* [fn. 69], 454 (incl. further ref.).
74 BVerfG, NJW 1977, 751 (751).
75 The ECtHR sees a positive obligation of member states only with regard to the creation of effective mechanisms for the protection of authors and journalists to allow for a public discourse, see e.g. ECtHR, 14.09.2010 - 2668/07, 6102/08, 30079/08, 7072/09, 7124/09, m.a. 137 - Dink v. Turkey,
76 See e.g. ECtHR, 26.02.2002 - 29271/95 m.a. 39 - Dichand and others v. Austria.
77 *Löber/Roßnagel*, [fn. 71], 151.
78 *Stark et al*, 2020, 31.
79 *Hindelang*, 2019, 181.
80 *Blake-Turner*, 2020, 1.
81 *Chambers*, 2021, 147 (3)
82 *Pearson*, 2021, 1181, cited after *McKay/Tenove*, 2020, 1 (4).
83 *Brown*, 2018, 194 (22)
84 *Stark et al*, 2020, 21.
85 *Kind et. al*, 2017, 4.

Electoral Freedom

Furthermore, an impairment of the free individual political will formation may also result in an impairment of the electoral freedom pursuant to Art. 38 Sect. 1 S. 1 GG. Its purpose is to ensure that voters make their (individual) choices in elections in a free and open process of opinion formation and express this undisturbed with the act of voting.[86] To reach this goal, Art. 38 Sect. 1 S. 1 GG warrants protection against compulsion and pressure and by doing so, safeguards against any serious influences by government and non-government players that might impair free decision-making.[87] In the opinion of the German Constitutional Court, the freedom of electoral choice also includes the option for voters to obtain information free of manipulation about election candidates. It considers deception and disinformation forms of influence that have the capability of seriously impairing the freedom to make electoral decisions[88], given that no serious defense option exists in this context.[89]

Hence, relevancy defined as capability to seriously impact the freedom of decision-making is required. It cannot be ruled out that disinformation reaches this level of quality, however, this is likely only the case in rare situations. Such a special risk can be seen in the (targeted) distribution of disinformation in the immediate time proximity of an election act. If the disposition of (targeted) disinformation occurs in this time window it is very difficult for discursive self-regulating forces in society to correct such an influence (see Chapter 3.3.4).

However, it once again has to be considered that the effects of disinformation and in particular its impact on the process of political will formation have only been inadequately examined empirically to date (see above).[90] There are also indications that the voter preferences are usually relatively stable despite election campaigns.[91] Moreover, influencing the election results is often not the immediate intention of disinformation campaigns, but rather more long-term shifts in discourses and discourse boundaries, social division or an erosion of trust in democratic institutions.[92] Moreover, a study conducted in the U.S. context indicates that voters who have not decided on their vote a few months prior to election day, tend to be less likely to believe ideological disinformation than voters who have already decided.[93] This is on par with the finding that individuals tend to specifically believe information that is congruent with their own world view (confirmation bias[94], see Chapters 2.2.4 and 4.3.1.2).[95]

Freedom of Expression

If the potentially misleading effects of disinformation put the autonomy and freedom of the individual (political) formation of will at risk, this also results in a risk potential for the fundamentals of opinion formation. Here, especially the effects on the observable opinion climate as a result of the use of social bots and its respective threat to knowledge may pose relevant risks. Art. 10 ECHR, Art. 11 CFR as well as Art. 5 Sect. 1 S.1 GG warrant the freedom of expression and the dissemination of one's opinion. A unique feature of German constitutional law is the non-inclusion of false claims of facts into the protective scope of freedom of expression (see above).[96] The consequence of this circumstance is that measures against untrue factual assertions regularly do not constitute an interference with freedom of expression. However, this only applies if there is no aspect of an opinion (e.g. statistical survey)[97], the person making the statement knows the untruth or the untruth is proven at the time of the statement.[98] Art. 10 ECHR and Art. 11 CFR on the other hand do not know similar restrictions of their scope.[99]

Disinformation does not directly impair the freedom to *express* an opinion, however, social bots distributing the respective information may lead to a change in the perceptible opinion climate[100] (see above Tab. 1). The false perception of the

86 BVerfG, NJW 1989, 1347 (1347) with ref. to BVerfG NJW, 1977, 1054.
87 BVerfG, NJW 1984, 2201 (2202) with ref. to BVerfGE 7, 63 (69); 15, 165 (166); 47, 253 (282).
88 BVerfG, NJW 1984, 2201 (2202) with ref. to BVerfGE 40, 11 (41).
89 BVerwG, NVwZ 2003, 983 (984 f.)
90 *Stark et al*, 2020, 31.
91 *Kalla/Broockman*, 2018, 148 (pp. 28).
92 *McKay/Tenove*, 2020, 1.
93 Indifferent voters are those who have not made their decision less than 3 months prior to election day: *Allcott/Gentzkow*, 2017, 211 (230).
94 See *Nickerson*, 1998, 175
95 Guess et al as well as Grinberg et al used sharing of disinformation as a proxy: *Guess et al*, [fn. 50], 1 (2); *Grinberg et al*, [fn. 50], 374 (5); *Hameleers et al*, [fn. 50], 281 (298); *Zollo*, 2019, 1
96 For instance, see BVerfG, NJW 1999, 1322 (1324) with ref. to BVerfGE 97, 391 (403).
97 *Schemmer*, 2019, m.n. 6 referencing BVerfG, NJW 1984, 419.
98 BVerfG, NJW 1999, 1322 (1324).
99 Cornils, 2018 , Art 10 m.a. 15; Cornils, 2018, Art.11 m.a. 32.
100 The term opinion climate describes the opinion of others perceived by recipients with regard to certain topics; cf. *Scherer*, 1990, pp. 19.

dominating relevant points of view in society[101] might, as a consequence, lead to a situation where individuals hesitate to express any deviating opinions.[102] It is presumed that people frequently base their opinions on the acts of others and fear the consequences of objecting to the (alleged) majority opinion; this is possibly also a reason to join the (presumed) majority opinion.[103] Hence, is some is led to believe in the relevance of untrue content, this could indirectly influence the freedom of expressing an opinion. A simulation study in this context showed that even if malicious bots make up only 2 to 4 percent of the opinions expressed on a platform, the opinion climate can be influenced in their favor in two thirds of all cases.[104]

Freedom of Information

Information plays an important role with regard to the individual's formation of opinions and the participation in the process of public opinion formation.[105] Art. 10 ECHR and Art. 11 CFR warrant the freedom to distribute and share information (active freedom of information), and also to receive information (passive freedom of information). Passive freedom of information is limited to generally accessible information.[106] Art. 5 Sect. 1 S. 1 Var. 2 GG also protects the negative dimension of freedom of information in the form of protection from imposed information.[107]

The positive dimension of freedom of information guarantees the individual's ability to access all generally accessible sources of information without hindrance.[108] Principally, the distribution of disinformation does not hamper the access to information, so that an impairment of the freedom of information can regularly be ruled out. However, things could be different if disinformation, through the artificial creation of relevance and/or reach reaches a level that factually rules out the visibility and access to other information or opinions. The threshold to identify such case as an impairment of freedom of information is relatively high, especially in media environments with an unlimited amount of information providers (see Chapter 2.1.3 and 3.3.2). In individual cases, the massive and vast distribution of false information with the objective of "drowning out" statements from trustworthy sources or traditional media may show a relevant risk potential.[109]

Beyond that, the negative dimension of freedom of information provides protection against undesired, imposed information.[110] Traditionally, it provides protection against government indoctrination (e.g. propaganda). Moreover, negative freedom of information becomes relevant when it comes to the unsolicited sending of information.[111] However, this cannot be understood as protection against any type of confrontation with information, which is why a certain significance must exist for the presumption of an impairment.[112] Excess as a result of the distribution of disinformation cannot be ruled out, but from a regulatory perspective, this comes with restrictive requirements, especially when it comes to political content.[113]

General Right to Personality

Besides the above-described autonomy aspects, the general right to one's personality pursuant to Art 2 Sect. 1 in combination with 1 Sect. 1 GG comprises a right to protection of social recognition and personal honor.[114] Disinformation may also contain statements that are libellous or contain denigrations of individuals or certain groups - either directly and unmistakably or through the creation of certain perceptions at the recipients' end. The distribution of untrue information is one communicative option among many aiming at the violation of personality rights. Targeted disinformation campaigns frequently aim at slandering certain groups or individuals through defamatory statements.[115] In this context, attention must be paid to the intersection of disinformation and the phenomenon of hate speech, given that potential legal rights

101 To be concise, mirroring of majority opinions in society is a specific function of media that is safeguarded by the objective dimension of media freedoms, see Chapter 2.2.4.
102 *Löber/Roßnagel*, [fn. 71], 154.
103 *Löber/Roßnagel*, ibid., 154 (incl. further ref.).
104 *Ross et al*, 2019, 394.; cf. the study on basis of 4.4 million tweets by *Kušen/Strembeck*, 2018, 37; *Kušen/Strembeck*, 2019, pp. 1.
105 BVerfG, NJW 2002, 2621 (2623); 2002, 2626 (2629).
106 See ECtHR, 26.3.1987 – 9248/81 m.a. 74 – Leander/Schweden; 16.12.2008 – 23883/06 m.a. 41 – Khurshid Mustafa and Tarzibachi/Schweden.
107 Grabenwarter [fn. 55] Art. 5 Sect. 1, Sect. 2 GG m.a. 1018-1019.
108 Cf. *Grabenwarter*, ibid. m.a. 996 (incl. further ref.).
109 Such practices are part of disinformation strategies. Steve Bannon called this tactic "Flooding the zone (with shit)", another term is "firehose of falsehood". Cf. Lewis, „Has Anyone Seen the President?", Bloomberg Opinion v. 09.02.2018, https://www.bloomberg.com/opinion/articles/2018-02-09/has-anyone-seen-the-president.
110 Grabenwarter [fn. 55] m.a. 1018-1019.
111 Grabenwarter ibid. m.a. 1018-1019.
112 Kühling [fn. 96] m.a. 44.
113 One example for a feasible regulation of imposed information in the commercial sector can be found in case law regarding unacceptable nuisances pursuant to § 7 UWG, s. *Köhler*, 2017, 253; *Scherer*, 2017, 891; see Chapter 3.1.1.
114 Di Fabio [fn. 55] m.a. 169.
115 *McKay/Tenove*, 2020, 1 (6); the authors explain this based on the example of Russian disinformation campaigns.

violations may not only be triggered specifically by untrue content, but by the contained violation of honor in the statement. Whenever untrue claims clearly violate personality rights, the government has the obligation to warrant an effective legal framework (see below Chapter 3.1.1).

Right to Unimpaired Personality Development (Protection of Minors against Harmful Media)

In Article 10 ECHR, the protection of children and adolescents primarily serves as a justification for interference with freedoms of communication, while Article 24 (1) CFR designates a much more comprehensive and much more positive aspect of the protection of children as a vulnerable group. The understanding of the constitutional protection of minors against harmful media goes even further, as stated in Art. 2 Sect. 1 in combination with Art. 1 Sect. 1 GG. Accordingly, the constitution gives rise to a state-directed duty to protect, which is aimed at ensuring an unimpaired personal development of minors. As normative development goals of children and adolescents, the BVerfG mentions personal responsibility and social skills.[116] Disinformation-related phenomena can impair the achievement of these development goals if, by believing respective (false) information, certain opinions or world views are established in such a way that they can have a negative impact on the social and ethical orientation of children and adolescents. In this respect, disinformation also shows risk potential for unimpaired personality development. An example of this are false information that fuels negative personal, gender-related or group-related prejudices or stereotypes (e.g. in relation to migrants, certain political parties or women). Furthermore, conspiracy narratives can also be relevant from the perspective of protecting minors.[117]

Body, Life and Health

The right to life and physical integrity is established by Art. 2 and 3 GRCh and partly in Art. 2 ECHR as well as by Art. 2 sect. 2 p. 1 GG. A danger to life and limb can arise if disinformation causes people to take further actions (or omissions), for example through information on how to deal with Covid-19 ("drink bleach") and measures recommended therein that lead to (self-)injuries. Inciting a mob with information about any wrongdoing by a particular person can also lead to violent riots in which people are ultimately injured or lose their lives.[118] In these cases, disinformation regularly does not represent a direct, but rather an indirect danger to life and limb.

The two aspects that must be considered here are, on the one hand, the causality between specific false information and concrete action, which is often not easily detectable[119]. On the other hand, endangerment of life and limb, especially in the case of acts of violence, are sanctioned by the legal system regardless of their cause (here: disinformation as a possible basis). The disinformation-specific risk potential usually can be seen here, too, in influencing the decision-making process of individuals. In blatant cases, however, the relationship between information and a concrete, immediate danger can be so close that regulatory intervention may appear possible or even necessary.

2.2.3. Group-related Risk Potentials

In addition to individual rights-related threats, disinformation can also show risk potential with regard to group-related legally protected rights. The prohibition of discrimination can be seen as a starting point regarding content that can lead to unequal treatment at the expense of vulnerable groups, especially those protected by law. In addition, group-related claims to social recognition that arise from the sum of the group participants concerned.

Both Article 14 ECHR and Article 21 CFR contain general prohibitions of discrimination. In German law, this derives from Art. 3 sect. 1-3 GG. Unequal treatment is only permissible if the differentiation criterion is suitable, necessary and appropriate with regard to the aim of differentiation.[120] Prohibitions of discrimination enshrined in fundamental rights primarily provide direct protection against state interventions and decisions of discrimination. Hence, prohibitions of discrimination

116 BVerfGE 121, 69 (92).
117 *KJM*, 2020.
118 This is not a purely theoretical possibility; in India, for example, several riots have been attributed to disinformation, cf. https://www.washingtonpost.com/politics/2020/02/21/how-misinformation-whatsapp-led-deathly-mob-lynching-india/; individual German studies also point to a connection between disinformation and acts of violence: *Müller/Schwarz*, 2017, 1; *Müller/Schwarz*, 2018, 1. These studies have been criticised in part , s. Cottee, 2018.
119 See *Bayer et al*, [fn. 13], 47.
120 *Kischel 2020*, m.a. 24 (incl. further ref.).

only offer a legal starting point if the widespread disinformation about certain groups is also reflected in discriminatory political decisions and state measures or if disinformation within the communication process leads to unequal opportunities for participation. The latter can arise, for example, when information and opinions of a certain social group are amplified to such an extent that minority opinions are severely limited in their visibility or their perceptibility is completely excluded. However, this mainly affects the more specific concept of a society-related right to equal opportunities in communication (see Chapter 2.2.4.).

2.2.4. Society-related Risk Potentials

Most frequently mentioned in the literature are risk potentials of disinformation for socially related legally protected positions and interests, and in particular the threat to public and political decision-making as well as the integrity of elections. In this context, it is sometimes even feared that democracy will be endangered or "destroyed" by disinformation. However, sufficient justifications and empirical evidence for these assumptions are usually lacking. Indirectly, there are also threats to public safety and order and, in certain areas, to public health.

Freedom of public opinion formation

The freedom of public opinion formation is the ratio legis[121] as well as positive legal obligation of the communication-related freedoms in Art. 5 sect. 1 GG and constitutes a fundamental condition of the principle of democracy laid down in Article 20 of the Constitution[122]. Similarly, according to the case law of the ECtHR, Article 10 ECHR also provides special protection with regard to the democratic will formation[123], whereby interventions with regard to public and political statements require a special justification.[124]

The freedom of individual and collective formation of opinion is constantly interrelated.[125] Potentially misleading effects of disinformation can thus lead not only to an impairment of individual opinion formation, but also to the freedom of public opinion formation with increasing exposure. In particular, the use of technical means to spread disinformation can intensify this risk. So-called "malicious bots" can be used for targeted communication of disinformation in large quantities (see Chapter 2.1.1).[126] This makes it possible to feign a seemingly high popularity for topics or persons[127], as well as to artificially create a high reach and relevance of disinformation. In addition, hashtags can be used to strategically amplify misinformation. For this purpose, often highly interconnected groups work together to make a hashtag a trend or to take over a currently trending hashtag.[128] In addition, topics that have a high relevance in social networks are often also adopted by traditional media[129] and sometimes even taken up by politicians.[130]

Through the massive distribution of disinformation, a distorted picture of the relevance of certain topics in public discourse can be created, which is not in line with actual discourse.[131] This can pose a threat to the free formation of public opinion, as a people's will is based to a large extent on an open and objective discourse[132]. A particular danger arises if an actually non-existing support for untrue facts is feigned, as this can additionally endanger the knowledge base for public opinion and decision-making (see Chapter 2.2.2).

The same possibilities for technical distortion and generation of attention can also cause a change in the climate of opinion. On the one hand, the massive spread of disinformation can lead to an overestimation of the weight of radical points of view

121 *Grabenwarter* [fn. 56)], m.a. 75.
122 BVerfG, NVwZ 2006, 201 (203).
123 ECtHR, 13.07.2012 - 16354/06 m.a. 48 - Mouvement Raëlien Suisse/Schweiz: "freedom of expression constitutes one of the essential foundations of a democratic society".
124 See e.g. ECtHR, 26.02.2002 - 29271/95 m.a. 39 - Dichand and others v. Austria.
125 *Hartl*, 2017, 30 (incl. further ref.).
126 *Löber/Roßnagel*, [fn. 72], 154; *Löber/Roßnagel*, 2019, 493 (494).
127 *Thieltges/Hegelich*, 2017, 493 (495).
128 *Agrawal*, 2020, 39.
129 *Kind et al*, 2017, 4.
130 *Ungern-Sternberg*, 2019, 13 (incl. further ref.).
131 See also *Steinebach u. a.*, 2020, 153. The author uses the term „distortion of public and political opinion"; it is stated: „A threat to the formation of public opinion can therefore be assumed in particular if disinformation is disseminated en masse. The more it is spread, the more it is likely to distort public and political opinion."
132 *Monsees*, 2020, 1 (122) (incl. further ref.).

in the discourse.[133] This in turn could increase the influence effects on users through disinformation (see Chapter 2.2.2). It should be taken into account, however, that there is evidence that the spread of disinformation continues to take place predominantly by humans, not by social bots.[134]

Equal Chance to Communicate

If the truth of a statement is doubted, at best a discourse at eye level leads to society discussing the truthfulness of the statement, thus constructing truth or untruth together and in exchange. In the ideal situation, a self-regulation of discourse takes place through speech and counter-speech in order to weed out the dubious and negotiate socially shared truths. However, the communication-related possibilities in digital networks make it possible to give disinformation (artificially) a higher reach and a supposedly greater relevance (see chapter 2.1.1). This raises the question of the extent to which the principle of democracy requires a process of negotiation and thus democratic decision-making in a fair way. In German constitutional law literature, the concept of equal chance to communicate ("kommunikative Chancengerechtigkeit") has evolved: This concept requires the open and privilege-hostile conception of communication processes[135], thus wanting to ensure the goal of diversity of opinion[136] and establish a condition for securing the free formation of public opinion.[137] However, if individual players in the public discourse gain positions of power that are not justified in a communicative way, a distortion of the public discourse takes place and this becomes "unfree" (see the remarks on the formation of public opinion below).[138] A similar concept has not yet emerged at European level, but comparable democratic premises can be found in the objective of equal access to information and freedom of expression in Article 10 ECHR and Article 11 CFR, as well as in the prohibition of discrimination under Article 14 ECHR and Article 21 CFR.[139]

Traditionally, the concept of equal chance to communicate aims at equal factual access to information sources and dissemination possibilities. However, equality in the relationship between individual statements is also to be achieved, whereby protection against systematic discrimination of certain points of view and their access to social discourse is to be ensured.[140] In order to enable such equal participation in the public opinion-forming process, it must be possible for everyone to participate in this process as a recipient or communicator.[141]

Applied to today's communication environment, the creation of artificial relevance and/or reach (e.g. through automated account networks, see Chapter 2.1.1) for the massive (targeted) dissemination of statements could therefore run counter to the ideal of equal chances to communicate. By doing so, a position of power can be taken in the discourse that is not justified communicatively but is due to the abuse of surreptitious communicative strength. Because the artificial increase of one's own visibility is accompanied by a lower visibility of other points of view, whereby legitimate voices in the public discourse could receive less importance[142]. This behavior is also called „inauthentic behaviour"[143], „inauthentic activitiy" oder „inauthentic engagements"[144] by platforms (see Chapter 4.4.1.).

At the same time, the attention logic of the platforms can also be exploited in order to achieve a particularly rapid and far-reaching spread of disinformation. Particularly exciting or emotionalizing (false) information usually spreads faster than true content, in particular when it concerns political and thus particularly important information for public discourse.[145] This proliferation of disinformation continues to take place predominantly by humans.[146] In view of the fact that the opportunity-oriented design of the communicative process is to be ensured, however, it is difficult to identify risks for equal chances to communicate, especially with regard to such organic dissemination. Because a position of power cannot be assumed solely because disinformation specifically arouses the interest of the recipients.

133 *Stark et al*, 2020, 3.
134 *Vosoughi et al*, 2018, 1146 (5). The study looked at particularly highly disseminated individual pieces of information (126,000 cascades of false or half-true information; most of the dissemination was through human or organic redistribution.
135 *Hoffmann-Riem*, 2010, pp. 668: cited after *Hartl*, [fn. 119], 32.
136 *Hartl*, ibid., 36
137 BT-Drs. 17/12542, 24.
138 BT-Drs. 17/12542, 24.
139 *Heldt et al*, 2021, 14.
140 *Heldt et al*, 2021, 14.
141 *Hartl*, [fn. 119], 32.
142 *McKay/Tenove*, 2020, 1 (5).
143 Facebook Community Standards (2021), https://www.facebook.com/communitystandards/inauthentic_behavior; Tiktok Community Guidelines (2020): https://www.tiktok.com/community-guidelines?lang=en#37.
144 Twitter Platform manipulation and spam policy (2020), https://help.twitter.com/en/rules-and-policies/platform-manipulation.
145 *Vosoughi et al*, [fn. 23], 1146 (2).
146 *Vosoughi et al*, ibid., 1146 (5).

Diversity of Opinions

In addition, amplifications of disinformation, i.e. artificially creating (alleged) interactions of users with a specific content, can represent a potential risk in terms of ensuring diversity of opinions. Diversity of opinion is a basic condition of the freedom of public opinion formation and is the goal of a communication process that provides equal chances to participate (see Chapter 2.2.3).[147] It is not sufficient for the goal of a self-regulating social discourse that there is just the possibility for participating for all points of view in the public discourse, but their *actual participation* is necessary. An intellectual debate and deliberation by society is made possible precisely by the actual participation of the diverse perspectives existing in society.[148] For this reason, diversity of opinion is a fundamental principle of freedom of expression under both European and German law. Ensuring diversity of opinion serves to protect the public communication process from influences by powerful actors.[149] Accordingly, pluralism also constitutes a positive legal obligations in Article 11 CFR[150] and Article 10 ECHR. The ECtHR stated that diversity of opinion is a constitute of a democracy[151] and that convention members have a positive obligation to ensure effective diversity.[152]

By exploiting the platform modalities and/or creating artificial relevance and scope of disinformation, and with it the distortion of discussion, positions of power of individual actors may arise that impair the perceptibility of other communication participants. This creates the risk that not all perspectives can form an equal basis for social debate.

Democratic Will Formation, Integrity and Electoral Freedom

The described threats of disinformation to the individual and collective formation of a political opinion also carry potential risks for the integrity of elections. Democratic processes are based to a large extent on open and objective discourse.[153] If the public image is distorted by disinformation with regard to the relevance of certain topics and/or individual (targeted) influences on the formation of a political will take place through disinformation, this could result in changed electoral decisions and, as a result, possibly even lead to a changed election result. Such influences by disinformation can constitute an impairment of the principle of free elections or electoral freedom pursuant to Art. 38 GG at the societal level.[154] The electoral freedom is intended to ensure on an individual level that voters can make their choice in a free and open process of opinion-forming and express it unimpaired in the context of the electoral act (see Chapter 2.2.2.).[155] However, if a considerable amount of individual voting acts is influenced by disinformation, this can be reflected in an overall altered election result and thus potentially endanger the freedom and integrity of the election, where the actual will of the people is to be formed by elections as an expression of popular sovereignty (Art. 20 sect. 2 GG). A correspondingly relevant influence through disinformation would thus run counter to the meaning and purpose of elections (see Art. 20 Sect. 2 GG).

Consequently, there are many concerns in the literature in this context, especially against the background of the increased use of targeted disinformation campaigns in the 2016 US election or the Brexit referendum.[156] A particular risk can be seen in disinformation activities in the immediate vicinity of the electoral act. For if the disposition to (targeted) disinformation takes place in great temporal proximity to the electoral act, any influence by it can hardly be corrected by discursive self-regulation of society. This idea is also expressed, for example, in § 32 BWahlG (Federal Electoral Law), which offers protection against influences in the immediate spatial and temporal proximity to the electoral act (see Chapter 3.1.1).

However, it must also be noted in this context that the effects of disinformation and in particular its influence on decision-making processes have so far been insufficiently investigated empirically (see above).[157] In addition, evidence on the stability of voter preferences[158], as well as on the likelihood of adoption of ideological disinformation by indecisive voters[159],

147 *Hartl*, [fn. 119], 36
148 *Heldt et al*, 2021, 15.
149 *Hartl*, [fn. 119], 35 with reference to BVerfGE 57, 295 (322); *Löber/Roßnagel*, [fn. 71], 152.
150 *Heidtke*, 2020, pp. 158.
151 ECtHR, 07.06.2012 - 38433/09 m.a. 129-130 - Centro Europa 7 S.r.l. and Di Stefano v. Italy; the court stated in the decision that there can be no democracy without pluralism.
152 ECHR, 07.06.2012 - 38433/09 m.a. 134 - Centro Europa 7 P.r.l. and Di Stefano v. Italy.
153 *Monsees*, 2020, 1 (122) (incl. further ref.).
154 The Federal Administrative Court (BVerwG) also considers deceptions and disinformation as influences which may be suitable to seriously impair the freedom of choice with regard to elections, cf. NVwZ 2003, 983 (984).
155 BVerfG, NJW 1989, 1347 (1347) with reference to BVerfG NJW, 1977, 1054.
156 *Marret*, 2020, 1 (3) (incl. further ref.); *Katsirea*, 2018, 159 (10).
157 *Stark et al*, 2020, 31.
158 *Kalla/Broockman*, 2018, 148 (pp. 28).
159 Undecided voters here are those who are undecided with less than 3 months to go until election day: *Allcott/Gentzkow*, 2017, 211 (230).

should be considered (see Chapter 2.2.2). In addition, influencing the election result is often not the main intention of disinformation campaigns, but rather the long-term social division and erosion of trust with regard to democratic institutions (see below).[160]

Trust and Democratic Institutions

The spread of disinformation could result in an erosion of trust in existing institutions;[161] this is also often the main goal of targeted disinformation campaigns.[162] Here, legal risk potentials arise on the one hand with regard to the principle of democracy and, sometimes, indirectly with regard to public safety and order (see below).

Democratic institutions are a basic condition of a functioning democracy. Through elections as an expression of a people's sovereignty, the legislature and indirectly the executive and judiciary branches are directly legitimized by the people, Art. 20 sect. 2 GG. The separation of powers serves to democratically balance by means of mutual control of the executive, legislative and judicial branches.[163] Accordingly, trust in these players is one of the basic conditions of a functioning democracy. If trust in these institutions is undermined by the influence of disinformation, this could potentially have consequences for the functioning of a democratic state, especially with regard to the decisions of the respective institution.[164]

However, an impairment of the functioning of democracy is highly demanding. Such an intensive and far-reaching influence on the population by disinformation cannot be assumed against the background of the limited knowledge of the effect in cross-media information and media repertoires, at least at this point in time.

Societal Construction of Reality and Social Cohesion

One of the prerequisites for living together in a society is the reference to a shared reality. In fundamental debates about disinformation, this aspect is emphasized and a necessity for action is derived from it. The importance of knowledge construction for living together is shared by the currently influential social theories, even if there are differences in the description of the construction of such a knowledge base.[165]

Beyond one's own world of experience, the construction of the world's knowledge is dependent on communication.[166] Society has developed practices, including building trust in functional elements of society that have recognized methods of testing truth, such as courts[167] and science.[168] These methods are fallible, but nevertheless they are regarded as useful possibilities for approaching the truth. If this trust is eroded, for example by current political and strategic claims in the United States that the presidential election was rigged, even though courts have rejected it, this has an impact on the democratic self-understanding of a society.

This effect can be partly caused and amplified by the dissemination of polarizing information. The spread of disinformation in particular is often attributed a polarizing effect.[169] Such polarization can be a catalyst for radicalization processes and could subsequently lead to a fragmentation of society. The problem of divergent constructions of reality within a society could be further intensified. Conversely, divergent reality constructions could in turn intensify polarization and fragmentation processes.

A basic question here is to what extent these prerequisites of a shared construction of reality can be seen as one of the preconditions that constitute a constitutional democracy, but which it cannot secure itself. Moreover, it is not yet clarified whether or to what extent – based on the respective constitutional system – the state is entitled or even obliged to counterfactually secure these prerequisites – in Germany, for example, as part of the positive legal obligations of Article 5 sect. 1 paragraph 2 GG. In any case, this remains a theoretical undertaking in terms of constitutional medial law, since it is not yet clear how instruments of communication control could actually counteract this at all. In this respect, this task initially

160 *McKay/Tenove*, 2020, 1 (1).
161 *Morgan*, 2018, 39 (39).
162 *McKay/Tenove*, 2020, 1 (1).
163 *Wank*, 1991, 622 (624).
164 *Reuter et al*, 2019.
165 Cf. *Berger/Luckmann*, 2016; *Luhmann*, 1988; *Luhmann*, 2018, pp. 83.; *Elias*, 2001, 133, 176-177.
166 *Luhmann*, 1996, 17-18.
167 *Seifert*, 2013, 155-156.
168 *Oreskes*, 2019.
169 *Bader et al*, 2020, pp. 57.; "Conflict of interests between the people and the elite or locals and immigrants".

falls within the scope of safeguarding political culture. Political players should be aware of the fundamental danger that a change in the practices of reality construction entails.

Public Safety and Order and Public Health

Public safety and order as well as public health are legitimate public interests under both German and European laws, which can justify infringements of fundamental rights. According to German laws, threats to public safety and order also empower respective law enforcement authorities to enforce police measures. Based on the individual risks to life, limb and health, disinformation can also lead to threats to these legally protected rights on the societal level. Examples are disinformation on the fight against Covid-19 as well as on vaccinations. Furthermore, disinformation that leads, for example, to violent riots and threats to life and limb on an individual level, can endanger the interest in public peace, security and order at the societal level. In this context, too, attention must be paid to the causality between concrete disinformation and its effects on legally protected social assets, which is sometimes difficult to prove. Threats to public safety and order as well as to public health are usually only the indirect consequences of disinformation. The disinformation-specific risk potential here, too, lies first of all in influencing the decision-making process of a large number of individuals. In exceptional cases, however, the connectedness between disinformation and the threat to public interests may be so close that legal intervention appears as a legitimate step. This may be the case in particular with disinforming incitements or encouragements to public nuisance (e.g. general incitement to violence against certain groups of society or state institutions). In addition, the mass spread of disinformation, e.g. on vaccinations or other health protection measures, can also attain such a degree.

Miscellaneous

Indirect consequences, e.g. for public safety, can also result from the further processing of data collected on social media platforms that contain disinformation.[170] The presence of false information in platform data, for example, is regarded as one of the central challenges for its use by humanitarian and civil protection organizations.[171] In addition, for example, in the context of the use of search instructions from social networks, threats to individual rights may arise.[172]

2.3. Dimensions of a Risk-Potential-Based Approach to Disinformation

Disinformation as a communication-based phenomenon is complex. Previous approaches to the description, definition, demarcation and categorization follow a phenomenological consideration, the core of which is the untruthfulness of a statement and the intention of the expressing person to mislead. From a legal or regulatory perspective, these two criteria show considerable difficulties when being used as legal criteria (see Chapter 2.1.3). In addition, such approaches do not make it possible to pre-shape the sometimes complex balancing processes when it comes to legitimate proportionate legal measures as well as concrete individual case-by-case decisions. In other words, the communication-scientific categorization of disinformation is helpful for the description of certain phenomena of deviant communication and for the differentiation of related forms, but the classification as disinformation does not yet say anything about the possible legal consequences. In the following, therefore, we fan out the dimensions identified as relevant on the basis of the identified risk potentials into a table format (see Tab. 2). These dimensions of disinformation are intended to help in the further study to systematically determine and define the scope of existing and possible new regulations, to check the admissibility of regulations (in particular their proportionality), and thus to create a basis for the selection of adequate governance measures.

170 *Pörksen*, 2018.
171 *Castillo*, 2016, 112 (incl. further ref.).
172 *Castillo*, ibid., 112. In the aftermath of the bomb attacks in Boston 2013, two uninvolved spectators were pictured on the title page of the New York Post after they had incorrectly been identified as suspects on the Reddit platform.

Tab.2: Legally Relevant Dimensions of Disinformation

Dimension	Explanation	Categories / Options
Type of statement	Differentiation between statement of fact and statement of opinion (relevance for orientation/suitability of counter measures, relevancy for necessity to take counter action)	· Pure factual claim (no counter-evidence possible; strong counter-arguments; objectively falsifiable) · Pure expression of opinion · Mix, e.g. opinion with a factual core
Context of statement	The level of protection granted by the BVerfG in balancing tests changes depending on the significance of the statement for social, above all political discourse. Statements in the context of purely economic competition also enjoy protection, but can sometimes claim less strong counterweight in the case of considerations with the rights of third parties. [Relevance to necessity of countermeasures; relevance for proportionality of countermeasures]	· No references · Purely competition-related statement / no points of contact with politics · Mixed formats: Points of contact with economic competition and political discourse · Political / ideological general discourse · Political / ideological discourse during election campaign
Actor structure	The nature or organization of the expressing entity may result in different regulatory approaches; in the case of automated communication, the freedom of communication as a legal position may also be weakened due to the lack of a right holder that can be attributable. The incentives or motives for the statement can also allow an important indication of possible regulatory approaches. [Relevance for orientation/suitability of countermeasures; relevance for proportionality of countermeasures]	· Individual person · Group of persons · Loose network · Formal network · Formal organization · Bot networks, fake accounts
Motives for statement	[Relevance for orientation/suitability of countermeasures; relevance for proportionality of countermeasures]	· Financial / economic motives · Political motives · Trolling / causing unrest
Degree of public visibility	The hypothetical and actual dissemination of disinformation has an impact on its influence on the processes of individual and public opinion formation. The aspect also plays a role by which (also technical) means the reach of an expression is increased. Problem: The achievement of reach is often intended, but can also arise from algorithmic artifacts; in these cases, the amplification may not be covered by the intention. [Relevance to necessity of countermeasures; relevance for proportionality of countermeasures]	· Only visible to the person issuing s a statement · Private communication · Group-internal communication · Platform-wide communication · Cross-platform communication · Adoption by traditional media outlets

Dimension	Explanation	Categories / Options
Recognizable intent to mislead / to deceive	The degree of intent allows for a better assessment of the proportionality of government countermeasures. [reprehensibility of an action; relevance for proportionality of countermeasures]	· Unaware · Negligent · Tolerated · Intentional

Source: Own source.

2.3.1. Risk Potential for Legally Protected Rights

For the analysis of possible legal reactions, the identification of legally protected rights and values that are affected by a statement is necessary. Only if legal rights could be endangered will there be any scope for state intervention that themselves touch fundamental rights. In determining the threat to legal positions, the probability of the risk materializing and its severity have to be taken into account together. The following applies: The more likely and the more intensively a potential impairment of a fundamental right might occur by disinformation, the sooner a countermeasure appears proportionate.

Tab. 3: Matrix of Potentially Affected Legally Protected Rights

	Degree of immediacy of the threat to a legally protected right / probability of a violation of a legal interest		Intensity of impairment of legally protected rights		Importance of the legally protected rights concerned
	Indirect threat	Direct threat	Weak	Strong /core area affected	
Individual fundamental rights and freedoms					
Individual autonomy / freedom of choice					
Unimpaired individual political opinion formation					
Electoral freedom					
Freedom of expression					
Freedom of information					
General right to one's personality					
Unimpaired development of one's personality / protection of minors					
Life, limb, individual health					

	Degree of immediacy of the threat to a legally protected right / probability of a violation of a legal interest		Intensity of impairment of legally protected rights		Importance of the legally protected rights concerned
	Indirect threat	Direct threat	Weak	Strong /core area affected	
Supra-individual legal interests					
Freedom of the public opinion formation					
Equal chance to communicate					
Diversity of opinions					
Democratic decision-making / integrity of the elections					
Trust in democratic institutions					
Societal construction of reality					
Public safety and order					
Public health					

Source: Own source.

3. IDENTIFICATION OF LEGAL GAPS

Chapter 2.1 has shown that a phenomenon-oriented understanding of disinformation is neither clearly determinable from a legal point of view nor realistically formalisable for legal decisions. Chapter 2.2 accordingly showed that forms of disinformation based on our work definition can show risk potentials for a range of legally protected individual and supra-individual rights and interests as well as the achievement of social goals. Based on that overview, dimensions of legally relevant dimensions were developed (2.3).

Existing national and European laws, with individual standards as well as with entire areas of law, already aim to protect the mentioned legal positions, supported in part by co- and self-regulatory procedures, standards and institutions. Amendments and reforms currently planned or under discussion as well as new regulatory approaches can be observed, some of which may significantly expand the current legal framework. In order to develop appropriate countermeasures in Chapter 4, the existing and foreseeable legal framework must be marked out in order to identify legal gaps in protection or areas with weak protection. In the following section, an overview of the existing legal and regulatory provisions is given from the specific perspective of the protected rights and interests presented in Chapter 2.2. It may happen that a certain set of standards or a certain provision shows connections to several protection objectives (see Tab. 3). Such redundancies are kept accordingly short given the special perspective of this approach. It was chosen in order to be able to quickly identify those protected positions and interests that have not yet been or only to a reduced extent been encompassed by the legal framework (see Chapter 3.2). In the final section the study identifies the possibilities and limitations of state measures to close the identified protection gaps, which will become the guardrails for the development of appropriate governance countermeasures in Chapter 4.

3.1. CURRENT LEGAL FRAMEWORK FROM THE PERSPECTIVE OF LEGALLY PROTECTED OBJECTIVES

3.1.1. Legal Framework for the Protection of Individual-Related Legally Protected Rights Threatened by Disinformation

The protection of individual rights by the legal system is one of the components of fundamental rights frameworks and one result of the rule of law. Accordingly, the current legal framework comprehensively encloses unjustified violations of positions protected by fundamental law.

Protection of Autonomy

The protection of autonomy with regard to statements in media is limited to commercial or commercially motivated statements or acts of deception resulting in financial losses, as in the case of fraud (§ 263 StGB; Criminal Code). Restrictions on particularly intrusive commercial communications that impair the consumer's freedom of choice can be found in competition law (§ 3 sect. 2, § 4a sect. 1, § 5 sect. 1 UWG; Fair Trade Law). However, the injunctive relief claims cannot be asserted by the end user, but only by competitors, competition associations, qualified institutions in accordance with § 4 UKlaG (Injunctions Act) as well as by the chambers of industry and commerce. Also if commercial statements directly affect competitors, they can also assert injunctive relief claims (§ 4 UWG). The protection of the autonomy of end consumers is only one of the legal objectives, though. In addition, competitors and other market participants are to be protected and fair competition is to be ensured overall. Besides competition legislation, advertising law in the field of broadcasting and telemedia also provides that a deception of recipients and users of media is excluded by the fact that commercial communication is always recognizable as such (§ 6 sect. 1 TMG; Telemedia Act), the MStV also provides for a separation requirement for the field of telemedia (§ 22 sect. 1 MStV; Media State Treaty). Autonomy-related effects of algorithmic selection logics outside of advertising communication, on the other hand, have so far neither been clearly located in terms of fundamental law nor

systematically included in law.[173] The transparency-related provisions targeting media intermediaries in the MStV, on the other hand, have been created primarily against the background of possible negative effects on diversity through algorithmic content selection.[174]

Protection of one's right to personality

A special role in the protection of individual rights is played by laws for the protection of the general right of personality and its manifestations. Where personality rights are violated by disinformation, those affected are entitled to a number of legal possibilities for their defense. Thus, in the case of defamatory allegations, criminal law norms may be violated (§§ 185 et seq. StGB), against which those affected can defend themselves with a criminal charge. For faster deletion of such statements, the NetzDG (Network Enforcement Act) comes into play. In addition, statements about individual persons always constitute the processing of personal data, so that data protection regulations are applicable; if the person expressing his or her opinion is unable to rely the data processing on a legal basis in accordance with Article 6 GDPR, the data subject has a right to erasure (Art. 17 GDPR). In addition, in the event of data protection violations, the competent supervisory authority can take action in the form of remedial measures (Art. 58 sect. 2 GDPR). Civil law can also be asserted against the person expressing a violating statement by filing for injunctive relief under tort law (§§ 823 et seq., 1004 BGB; Civil Code).

The scope of the civil law injunctive relief claims goes beyond the restrictive scope of application of the criminal protection of honor. In any case, those affected can take civil action against unjustified reporting by journalistic media with regard to their own person in conjunction with § 22 f. KUG (Art Copyright Act). The extent to which claims from the KUG against unjustified publications about one's own person still exist in addition to the GDPR framework outside of journalistic reporting is seen as controversial. Against factual claims concerning him or her in journalistic-editorial outlets, a person concerned may have a right to reply to be enforced under civil law pursuant to § 20 MStV, i.e. the expressing entity may be obliged to publish the affected persons own statement. The counter-statement, which is to be directly linked to the initially published information, may contain only factual assertions.

In media law, § 19 sect. 1 MStV the requirement for telemedia with journalistic content states that they "have to comply with the generally accepted journalistic principles". With the aim of protecting individual rights[175]– in particular the prevention of violations of personality rights – and to secure the process of public opinion formation[176], journalistic offers[177] must thus adhere to journalistic duties of care in their reporting. In particular, this includes the obligation to check the content, origin and truth of news before they are distributed with the care required under the circumstances. Sentence 2, which has been added to the new MStV since autumn 2020, extends these duty of care obligations to all business-related, journalistically and editorially designed telemedia services "in which news or political information are regularly contained". Now all statements which regularly contain current topics and are not provided purely privately have to adhere to appropriate duties of care.[178] The supervision of compliance with the obligations of these offers is carried out by the respective competent state media authority, insofar as the service provider is not subject to the German Press Code or has joined a (as yet non-existent) self-regulatory institution. In case of violations of § 19 Sect. 1 MStV, the responsible state media authority may initiate complaint proceedings against the provider (§ 109 MStV). With regard to alleged disinformation, § 19 sect. 1 p. 2 MStV is relevant: Here a supervisory body does not draw regulator measures on the truth or untruth of information, but rather on the compliance with structural journalistic measures to guarantee truth (see Chapter 3.3.4). If a content provider who has signed up to the Press Code or a self-regulatory body does not comply with the duty of care obligations, the Press Council or the self-regulatory body may, as a last resort, issue a public complaint with an obligation to publish it.

Protection of Minors

The right of the individual to an unimpaired personality development is, as described above, at the core of the constitutional duty to protect minors against harmful media content. Federal and state legislators have transposed this mandate in the JuSchG (Youth Protection Act) and the JMStV (Interstate Treaty on the Protection of Minors in the Media). Disinforming statements can have relevant aspects with regard to these legal frameworks. § 6 JMStV protects children and young people from advertising communication that exploits their inexperience and gullibility or that generally impairs them phys-

173 See *Dreyer/Heldt*, 2021, 117.
174 Cf. Official Justification § 93 MStV.
175 Cf. *Lent*, 2020, 593 (593).
176 Regarding the social aspects of duty of care obligations cf. 2016, pp. 359.
177 Regarding the scope of application cf. Lent, 2020, 593 (597).
178 Regarding the difficult demarcation cf. Fiedler, 2021, m.a. 8-10.

ically or mentally. Even outside of commercial communications, disinformation can have adverse effects from the point of view of the protection of minors, for example where world views and perspectives are propagated through untruths and half-truths that are detrimental to the developmental goals of personal responsibility and social skills. Statements with potentials for such detrimental effects can impair a minor's development within the meaning of § 5 sect. 1 JMStV and must be blocked by the content provider with appropriate access barriers.

Platforms frequently used by minors and who provide development-impairing disinformation are not themselves responsible for these statements as long as they are not aware of them. According to § 24a JuSchG, however, they are obliged to implement effective preventive measures in order to safeguard the goals of the protection of minors. The legal provisions in this area are primarily aimed at making it more difficult for children and young people to come into contact with relevant content. The regulations are enforced by the competent state media authorities and their central body KJM (Commission for Youth Protection) when it comes to JMStV provisions and, if necessary, by approved bodies for voluntary self-regulation. In the area of the infrastructure-related requirements of § 24a JuSchG, supervision is exercised by the new BzKJ (Federal Agency for the Protection of Children and Young Persons in the Media).

3.1.2. Legal Framework for the Protection of Group-related Legally Protected Rights Threatened by Disinformation

Legal regulations that protect social groups in their rights primarily concern protection against discrimination, with a focus on forms of group-related misanthropy.

Protection against group-related misanthropy

According to § 130 Sect. 2 No. 2 StGB the incitement to hatred spread via telemedia and the incitement to take violent or arbitrary measures against certain groups, parts of the population or individual members of a group or part of the population on the basis of their affiliation is punishable. Where definable groups of people are affected by attacks on their human dignity through insults, malicious slander or gossip, the prohibition of sedition opens up possibilities for protection. In contrast to statements made in front of those present, the fulfilment of the criminal provision does not require an aptitude to disrupt the public peace. Especially in the context of disinformation, however, a simply derogatory statement or a tendentious half-truth depiction is usually not sufficient to fulfill the facts, as § 130 StGB presupposes a particularly qualified form of impairment, namely "a violation by massive attacks of the persons concerned in their fundamental rights to life as equal personalities in the community, through which the indispensable area of the personality core receives a social devaluation".[179] For statements that, on the basis of untrue factual assertions, give a certain (negative) impression of a definable group of persons, the offence of sedition cannot therefore be used without further ado.

Protection against group-related discrimination

The prohibitions of discrimination established in the AGG (General Act on Equal Treatment) only apply in certain social contexts (§ 2 sect. 1 AGG) as well as in the establishing of private law obligations (§ 19 sect. 1 AGG). In this respect, devaluations of certain sections of the population contained in untrue allegations do not fall within the direct scope of these statutory provisions.

In the area of media law, there are also provisions that aim at enabling opinions and perspectives occurring in a society to be seen in public discourse, such as the diversity-related legal requirements for public service broadcasters (see § 26 MStV). However, there are no comparable requirements for private broadcasters and telemedia providers. Hence, corresponding demands for diversity cannot be converted into a claim by collectives against forms of content-based or technical discriminations. The concept of equal chances to communicate (see Chapter 2.2.3) is not yet reflected in individual norms of media law – at least not in relation to specific population groups.

179 *Schäfer*, 2017, m.a. 50.

3.1.3. Legal Framework for the Protection of Societal Goals Threatened by Disinformation

With regard to the legal protection of societal goals and collective rights, the relevant legal framework remains negligible. The risk potentials described above for the freedom of public communication and the associated possibility of democratic decision-making are legally addressed in very specific constellations only.

Protection of free public communication and democratic opinion formation

Examples of this are prohibitions on election interference as stated in § 108a StGB or § 32 BWahlG. The criminally relevant voter manipulation includes above all an influence on a person entitled to vote by deception with the result that the person concerned is wrong about the content of his vote decision, does not vote against his will or does invalidly vote. The aim is to protect not only the freedom of choice of the individual, but also the election as a whole against falsification.[180] To change someone's mind (also by deception) to vote for a different party than originally intended does not fall under the criminal offence. Against this background, statements that make false claims about the procedure of filling in or submitting ballots and thus, for example, cause voters not to submit or submit invalid ballot papers can be criminally relevant. § 32 BWahlG, too, aims at a prohibition of influence in connection with an electoral act. Accordingly, any form of influencing voters in direct temporal and spatial connection with an election is prohibited (Sec. 1). Since the publication of voter surveys even before the polling stations close can also have an impact on people who have not yet voted, the publication of corresponding preliminary results is forbidden (§ 32 Sec. 2 BWahlG) – even if they contain only supposed preliminary results.

Securing Media Diversity in Media Law

The legal provisions regarding media intermediaries introduced with the MStV contain a specific prohibition of discrimination through algorithmic prioritization (§ 94 MStV), which is directed against the unjustified discrimination or obstruction of journalistic offers; in essence, the provision is committed to ensuring diversity. Insofar as intermediaries reserve the right to take into account disputed statements or untrue factual assertions of journalistic-editorial telemedia less strongly or not at all when displaying search results or in recommendations, this circumstance cannot be assumed to be improper discrimination as long as they disclose this in the context of their transparency obligations pursuant to § 93 sect. 1 MStV. In this respect, § 94 MStV opens up the option for intermediary providers to take action against providers who have already spread disinformation several times by means of so-called depriorizations (see chapter 4.3 below). However, a legal obligation to act in this way does not result from the MStV.

Another new legal situation results from the transparency obligations that the MStV provides for automated accounts such as e.g. social bots. According to § 18 Sect. 3 MStV and § 93 Sect. 4 MStV both the content provider and the media intermediaries are obliged to mark automatically created content. The prerequisite is that the corresponding user account "has been made available for use by natural persons according to its external appearance". By marking the circumstance that an account is communicating on an automated basis, automated statements are clearly recognizable and can be critically reflected during the discussion process: Here, a non-human actor participates in the public discourse. For the area of automated accounts that disseminate or share dubious statements, the requirement of § 18 Sect. 3 MStV might be helpful insofar as it helps to reduce potential deceiving effects on the recipients' side.

Protection of Public Safety and Order

As shown above, it is conceivable to have situations in which social interests in the form of public safety and order may be indirectly affected by a statement, for example by the praise of certain behaviors for the alleged protection of health (drinking bleach) or the insinuation that individual persons or groups would enjoy special advantages at this moment. Behavior or outrage caused by this can then grow into direct threats to legally protected rights. In such cases, it is controversially discussed to what extent police authorities can take action against such media statements, for example in the form of interim injunctions within the framework of police law. The dispute is ignited by the fact that journalistic media are regularly "police-proof", i.e. that no police measures are allowed beyond the measures of media regulatory supervision; corresponding blocking regulations can be found regularly in the press and media laws[181], and the wording in § 17 p. 3 MStV is also understood in this direction.[182] However, with regard to media freedoms, it remains unclear to what extent police measures against non-journalistic telemedia are possible. As long as the offers concerned are relevant for the formation

180 *Eser*, 2019, m.a. 1.
181 Cf. § 1 (3) LPG NRW; § 1 (3) HmbPresseG.
182 *Möller et al*, 2020, 56.

of individual and public opinion, the same consideration of police proofness should be applied in principle[183]; only for statements that have no relevance for opinion-making police law could then be applied. As measures are based on the content of a statement and its resulting dangers, then in case of doubt, the idea of media freedoms blocking countermeasures applies to journalistic telemedia.

3.1.4. Latest Developments of the EU Level

At European level, there are two broad areas of activity related to the phenomenon of disinformation.

In the "European Democracy Action Plan", the EU Commission has made it clear that the decision on truth or untruth does not belong in the hands of the state ("No Ministry of Truth"). Overall, it aims to improve the coordination of activities of the EU and the Member States, in particular through the Rapid Alert System (RAS) and the European Cooperation Network on Elections (ECNE). These two cooperative initiatives have a strong focus on elections. In addition, it has been announced to develop guidelines for the duties and responsibilities of platforms in the fight against disinformation, with a special focus on the creation of transparency in political advertising (see Chapter 3.3.5). This includes testing ways of more effectively cracking down on foreign interference in elections.

A central project announced is also the revision of the "Code of Practice on Disinformation" from the current purely self-regulatory initiative towards a more co-regulatory approach. In a first step, a sustainable framework for the implementation and monitoring of the code is to be created. The Code of Practice and its planned modernization are at the heart of the EU's disinformation-related activities. As a standard, the code represents rules of conduct for the industry, which individual platform and content providers can join. So far, the code has 16 signatories. In recent months, the EU has initiated a special COVID-19 monitoring and reporting program in the context of the Code in order to better monitor disinformation in the context of the novel coronavirus. Following an evaluation in spring 2021, the EU Commission is currently putting together optimization proposals, of which the following key points are known[184]: The code is to be revised in such a way that the spread of disinformation via ads is reduced and that monetary incentives for distributors of disinformation are reduced. In addition, fact-checking initiatives are to be further supported, in particular in the design of transparent and non-discriminatory procedures as well as in the cooperation between fact checkers and the major platforms.

Another focus will be on improving the integrity of systems with which platforms detect and limit the artificial increase in reach. On the side of particularly desirable content, the code is to be developed in a direction that also stimulates an exchange of possible criteria for the identification of such socially valuable content. This requires studies on indicators for the trustworthiness of sources and offers as well as better access to data in the hand of the of the platform providers. As in many other policy areas, the EU also aims to build key performance indicators to better measure the effects of disinformation and government countermeasures. In the context of the EDAP, the European Digital Media Observatory (EDMO) has also been established, which is to act as a central point of contact for monitoring disinformation activities and for fact checking.

The draft Digital Services Act (DSA) focuses on transparency requirements, the guarantee of user rights, access to platform data and the handling of illegal content. Legal forms of disinformation are not directly covered by the draft. Nevertheless, there are some connections: for example, very large platform providers are to be obliged to carry out systemic risk assessments (Art. 26 DSA proposal). However, the relationship between systemic risks and disinformation remains unclear. For example, the misuse of platforms by fake accounts and bot networks would appear to be a systemic risk for a platform, but the extent to which the use of platforms for the "simple" publication of disinformation falls under it is rather questionable. Overall, it is still unclear to what extent the impact assessments can also include endemic risks, i.e. risks occurring outside the platform. Especially from the area of NGOs fighting disinformation, there are signals of disappointment. The focus on the large platforms is too strong; also and especially on smaller platforms there are sworn communities in whose context disinformation and conspiracy myths can arise. The DSA draft also lacks an obligation to verify accounts. Overall, disinformation appears primarily as a problem of content moderation in the draft, no instruments

183 See *Möller et. al*, ibid., 57.
184 European Commission Guidance on Strengthening the Code of Practice on Disinformation (COM (2021) 262 final), 26.05.2021, https://ec.europa.eu/newsroom/dae/redirection/document/76495

against cross-platform, long-term campaigns are provided, and the combined use of small websites and platforms as distribution channels is not taken into account either.

3.2. IDENTIFICATION OF GAPS IN PROTECTION

The contrasting of the rights and interests at risk (Chapter 2.2) with the current legal framework (Chapter 3.1) shows, using the example of Germany, that at least individual and group-related legally protected rights are in principle also protected against dangers emanating from (online) communication. Here, there is comparatively comprehensive protection for the legal rights of individual persons affected – especially where the general right of personality is concerned. In practice, this means that violations of rights in the intersection of disinformation and criminally relevant hate speech as well as violations of personal rights under civil law are already covered in a variety of ways. However, where disinformation does not violate individual rights and does not process personal data, legal countermeasures are limited. In these cases, at most, sector-specific provisions appear in special cases; these include provisions from the youth media protection law in cases of statements that have development-impairing effects for children and adolescents, or forms of misleading end users that may violate competition regulations. At the level of individual protection, it has become clear that, in particular, forms of specific information-based deception and deception without financial losses caused by this are not covered by the current legal framework. In particular, this can lead to open flanks in the protection of individual (political) decision-making processes.

There can be criticism of the effectiveness of the existing frameworks protecting individual rights with regard to the possibilities of the legal system to deal with the volumes and speed of distribution of potentially rights-damaging information on platforms. However, this is a problem of all impairing content and not limited to expressions that bear risks because of their possible untruth.

At the level of group-related protection of legal interests, the inventory has shown that instruments against disinformation-related risks only exist in a few areas – which are rather irrelevant from a disinformation point of view. In the area of criminal law, there is basic protection for at least the most massive violations of group-related legal positions. In practical terms, the specific element of disinformation is likely to recede into the background when it comes to statements with such forms of extreme contempt.

Content that triggers dangers because of its possible untruth also appears as an issue when it comes to the possibilities of platform-related legal reactions in order to make the protection more effective: If the infringement of legally protected rights lies precisely in the untruth of a claim about a person, for example, an adequate solution to a legal dispute presupposes that – depending on the burden of proof – the truth or untruth must be proven. Regulatory approaches such as those of the NetzDG (Network Enforcement Act) require the platforms to make this decision, even though they do not have court-like proceedings that can ascertain truth. Chapter 2.2 shows how demanding a procedure that preserves freedom of communication is – it is not only about the determination of the facts, but also about the question of whether there is an assertion of fact at all and what exact meaning it has. The incentives for platforms to delete doubtful content combined with an obligation to check the content lead to doubts as to whether such regulatory concepts are at all compliant with fundamental and human rights.[185]

With regard to societal interests, concrete regulations in Germany so far only aim at legal interests such as public peace and the counterfactual stabilization of trust in journalistic content (see Chapter 3.1.3). In addition to traditional restrictions on concrete actions influencing the electoral decision, there are also newer forms of media regulation, which are aimed at the non-discrimination of individual (journalistic) voices in intermediary services, as well as transparency requirements for forms of automated communication. At the European level, the planned activities are primarily aimed at political communication – in the form of rules for ads and micro targeting. The constitutionally "softer" – i.e. less tangible – goals, from which at best abstract demands for action arise for the state (democratic decision-making, trust in democratic institutions, social cohesion), are currently not found as concrete protective purposes of information-related requirements in the legal framework.[186]

185 *Claussen*, 2018, 110 (119).
186 Cf. *Buchheim*, 2020, 159 (166).

3.3. OPTIONS AND LIMITATIONS OF STATE MEASURES TO CLOSE LEGAL GAPS

Overall, the current legal framework has only very limited points of contact for curbing the specific risk potentials of disinformation, especially where statements do not (at the same time) violate individual personality rights. The notion of a protection gap implies that the state could be required to make immediate improvements and to systematically include the phenomenon of disinformation into the regulatory frameworks. However, the question of whether there is a need for action is first and foremost a political one, whereby the political scope for action has limits in both directions: On the one hand, constitutional frameworks can contain positive legal *obligations* for the protection of rights; however, with regard to legislators, these obligations are rarely concrete enough for certain legislative action to be imperative. On the other hand, human rights frameworks set limits to state action; in principle, a concrete assessment can only be carried out on the basis of concrete proposals for legal interventions. Therefore, in the following section, these limiting guidelines are outlined in order to narrow down the scope for possible political action. In Chapter 4, the study then focuses on possible countermeasures that are not questionable from the outset and the effectiveness of which seems at least plausible.

3.3.1. Limits of Truth/Untruth as Statutory Criteria

As shown above, the central starting point of most scientific definitions of disinformation is the untruth of statements (see Chapter 2.1). The wide variety of proposed countermeasures also regularly assumes that a content has been identified as untrue; the suggested measures just presuppose this. However, how the statement was identified as problematic and the falsehood was identified is not part of the corresponding works. Even legal papers skip this point and present the untruth as a given.[187] From a legal point of view, however, the determination of the truth of a statement is not only a complex challenge (see Chapter 2.1.2), which legal proceedings approach in individual cases through judicial assessment of evidence in court, it must also first be included in fundamental preliminary considerations that ask whether the law-making state meets the task of deciding on truth and untruth at all – or may do so.

A state deciding on what is true and what not is in conflict with the basic assumption that the negotiation of truth is a task that is in the hands of society itself. Truth is primarily negotiated and socially constructed through social discourse. Society also gives itself the rules of this negotiation process, which enables a shared basis of reality and thus societal knowledge, which is the basis for individual and collective action. The principle of democracy is based on the assumption that the formation of a people's will extends from the people through elections to the state. As the Federal Constitutional Court makes clear: the process of opinion formation should take place without state interference, the political decision-making process should take place from the people to the state, and not vice versa.[188] This basic principle also finds its counterpart in constitutional media law that requires limited influence on the state on broadcasters.[189] If the state intervenes directly or indirectly in the corresponding processes through legal interventions, there is always the danger that the state thereby reverses this principle and prescribes or at least suggests to society a pre-shaped truth. The risk of state abuse of interventions in public expressions cannot be denied with regard to authoritarian regimes.

It is merit and hope of the enlightenment[190] and also the basic assumption of freedom of expression that discourses are at best rational, constructive and understanding-oriented.[191] However, freedom of expression cannot guarantee this – from the point of view of an informed, deliberative discussion and the exchange of rational arguments, freedom of expression might even appear as a risk.[192] Ensuring "good discourse" through state intervention is neither desirable nor feasible; this refers to the Böckenförde dilemma: "The liberal secularized state lives by prerequisites which it cannot guarantee itself."[193] In the very moment that the state dictates what is true and what is false, freedom ends. For legislative activities in the area of the freedom of expression, this means that these can only be carried out selectively – if at all – and limited to specific hazards and taking into account the freedoms of communication and the principle of the rule of law. A possibility to prohibit or criminalize the spread of untruths, uncomfortable views or simply marginal ideas in general is excluded.[194]

187 Cf. *Mafi-Gudarzi*, 2019, 65.
188 BVerfGE 20, 56 (99).
189 Grabenwarter [fn. 55], m.a. 830 et passim.
190 *Habermas*, 1990, 182-183.
191 *Jestaedt*, 2011, m.a. 9., who points out, however, that the expectation that the truth will be ascertained is not to be equated with truth as an object of protection under Art. 5 sect. 1 GG .
192 Cf. *Masing*, 2012, 585 (585).
193 *Böckenförde*, 1991, 92 (112-113).
194 *Buchheim*, 2020, 159 (168-169).

Moreover, using the falsehood of a statement as a legal criterion is not possible at the level of facts in light of the entanglement of factual claims and the expression of opinion. Conversely, this does not mean that there is a "right to lie".[195] In individual cases, though, assertions objectively identified as false are subject to lower justification requirements than value judgments in the case of legal interventions.[196] If restrictive consequences (e.g. omission, deletion) are linked to the identification of the untruth of an assertion, uniform (court-like) case-by-case procedures are necessary, which work independently, contradictory and with the help of evidence gathering procedures. This has consequences for the scalability of such findings.

Statutory communication regulation is limited to narrow areas of criminally and personality rights relevant statements and depictions with immediate potential for danger. In determining such immediacy, recourse can be made to the legal concept of danger in the area of police law, according to which a danger is immediate when the damaging event has begun, or if this development is imminent immediately or in the near future with a probability bordering on certainty.[197] While the existing legal framework already pursues individual rights-related protection purposes (see Chapter 3.1.1), the question arises in particular as to legislative possibilities for safeguarding the basis of collective decision-making. This is especially relevant in cases where an information has relevance for the public formation of a (political) will and a societal negotiation of their truth is not possible for temporal, local or structural reasons. In constellations in which the public discourse regarding a statement and its truthfulness is not possible, the principle of social self-understanding tilts in the direction of a state that compensates for the impossible negotiation process in the event of a concrete threat to legally protected rights, applying an objective balance between the freedom of expression and the protection of collective decision-making bases, exceptionally deciding instead of society. Examples of such special circumstances that have come to light in this report mainly concern doubtful statements in the immediate vicinity of an election (see also Chapter 3.3.4).

For all other forms of statements, the truthfulness of which can be determined by social negotiation processes, the principle of state restraint applies. Here, society itself is foremost called upon to negotiate the truthfulness of statements.[198] However, where these negotiations are increasingly taking place in forums that are publicly accessible but organized under private law, the question arises as to whether and with what means the state can guarantee or at least support the truth-related negotiations. If the truth of a statement is disputed in the process of discussing and reasoning, it is first of all the goal of communicative exchange to make doubts visible – and at best to provide counter-arguments or counter-evidence that strengthen the counterpoint in a discursive way. Doubts indicate the denial of the claim to truth of a statement. Where governance measures aim at supporting social discourse for the negotiation of truth in spaces under private law, enabling forms of raising doubts appear central.

3.3.2. Positive Legal Obligations Enabling Equal Chances to Communicate

The state is largely denied legal intervention in relation to specific content of statements. At the same time, however, the state is subject to positive legal obligations[199], as far as the safeguarding of the basic conditions of a free public communication constitution[200] is concerned.[201] It has "to design the communication order in such a way that every empirical subject has the real chance to participate actively and passively in the communication process".[202] The aim of these efforts must be to maintain the conditions "that make the process of public communication, 'the constant intellectual debate, the struggle of opinions', possible in the first place."[203] If the state pursues these obligations, the focus of state control is not on individual statements, but on the "prerequisites for the creation and maintenance of the communication process in which every statement is embedded".[204] Thus, legal measures appear possible where the legislator can identify a threat to these conditions; here, the lawmaker has a margin of appreciation.

195 The law calls for exceptions from this in some cases in criminal law and with impermissible questions in applicant interviews.
196 *Jestaedt*, [fn. 35], m.a. 37.
197 Cf. the police laws of the German federal states.
198 *Schaal*, 2020, 347.
199 The question of positive legal obligations of Art. 5 Sect. 1 GG and its reach is a field of extended constitutional discussions, which also contain opinions that are critical of this point of view, see *Jestaedt*, [fn. 35], m.a. 18 et passim.
200 *Hoffmann-Riem*, 2001, m.a. 41.
201 On these structural guiding principles cf. in detail *Heldt et al*, 2021.
202 *Schulz*, 1998, 178.
203 *Grabenwarter*, [fn. 55] m.a. 109.
204 BVerfGE 97, 391 (399).

With regard to the dangers of legally protected social interests described in Chapter 2.2, corresponding measures could refer to situations in which certain (legal) views are systematically excluded from the discourse, or a certain view holds a communicatively unfounded supremacy through tools of (technical) assistance and thereby displaces other viewpoints. The former example has strong references to the diversity-related requirements of the positive dimension of the freedom of broadcasting, which have found their transposition in media concentration rules (§§ 59 ff. MStV) as well as positive diversity-related requirements (§§ 51 Abs. 2, 59 Abs. 1 MStV) in the MStV. Other relevant constellations can be seen in cases where privately owned platforms and intermediaries decide to delete certain opinions and views without any objective reason; in this respect, the prohibition of discrimination directed at so called media intermediaries also shows references to this policy area (see Chapter 3.1.3). The additional example – the obtaintment of communicative power through the exploitation or misuse of communication possibilities – can also trigger such positive obligations. The only legal provision in this direction so far is the transparency requirement for automated profiles in social media, which is aimed at profile owners and platforms (§§ 18 sect. 3, 93 sect. 4 MStV).

A legal restriction of larger campaigns, which obtain a communicative advantage through the purchase of followers, likes and (further) distributors, is not yet in place. Such measures seem possible at least with regard to cases of far-reaching, massive (dis)information campaigns. The main challenge here is to develop and implemtent concrete and comprehensible legal as wellas contractual criteria (see Chapter 4.4.1).

The problem of a non-communicative based position of power also arises in constellations in which the state gives itself the right to place particular emphasis on statements in public communication. Accordingly, forms of obliging private media actors to publish official statements are limited to public warnings in the event of disasters or situations of imminent danger only, focusing on warning the population. The condition of significant public health risks can also be met in rare cases by disinformation, for example in the area of the request to carry out actions that are directly self-endangering or pose a risk for pulic safety. If a supervisory procedure does not promise a sufficient remedy in terms of time, or if dangerous allegations are disseminated in many variants and massively, public announcements could also appear suitable in the field of telemedia in exceptional cases to counter these dangers.

3.3.3. Possible Legal Requirements Concerning Statements with a Special Claim to Truth

A particular challenge of the societal discourse-based negotiation of truth arises where special trust is placed in publicly available communication – for example, because the person or institution making a statement is ethically or legally obliged to tell the truth (e.g. public authorities[205]) or has to comply with professional duties of care, obliging them to report as truthfully as possible (esp. journalism[206]). Where such requirements exist for the communicator, this leads to an increased trust in the truthfulness of a statement on the part of the receiver – in the sense of correctness, completeness and topicality[207]. Here, scientific evidence is available that points to the special role of such content for publicly relevant communication. There is a social interest in securing these expectations, in particular with regard to the separation of opinion and factual assertion as well as the efforts of journalistic actors to ensure correctness. In media law there are corresponding legal requirements for compliance with journalistic principles and obligations to stick to professional duties of care[208] to ensure the journalistic fulfilment of democratic functions. The purpose of such requirements is the protection of personality rights of individuals against impairments of their social recognition, but also the interest of the general public in truthful reporting.[209]

Where content *imitates* journalistic statements without adhering to truthfulness-related duties of care, the question arises whether and to what extent the legislator is allowed to extend legally binding requirements to these actors. The purpose of such rules can be seen in the counterfactual stabilization of trust in journalistic content. This way, statements that assert an increased claim to truth through their style or design would be obliged to stick to higher requirements regarding the truth than non-journalistic expressions. § 19 Sect. 1 p. 2 MStV is, as shown above, an existing legal provision that obliges

205 On the truthfulness and correctness obligation for government communications *Mast*, 2020, pp. 244.
206 *Bentele*, 2021; *Stapf*, 2012.
207 *Pöttker*, 2017, 85.
208 See § 6 LPresseG NRW, § 31 Sect. 5 LMG NRW, § 19 MStV.
209 *Schierbaum*, [fn. 170], 259-260; *Heins/Lefeldt*, 2021, 126 (128).

providers of journalistically or editorially designed content to comply with appropriate duties of care. The provision – and generally forms of increased requirements for non-traditional, journalistic publications – meets constitutional concerns in literature, especially with regard to freedom of expression, insofar as they link a vague legal scope of application with increased requirements regarding the production of statements in publiccommunication.[210] On the other hand, a sliding scale of duties of care, which is oriented towards public communications and their relevance to the formation of opinions, can enable graduated requirements corresponding to their social responsibility.[211] From this point of view, graded forsm of dutiesof care could result along the criteria of an objective claim to truth (in particular through journalism-like design), broad impact (circle of recipients) and topicality. Because trust is constructed mainly on the part of the recipients, legal procedures are regularly limited to case-by-case decisions. In view of the potential for abuse of legal duty of care obligations, especially with regard legal actions against certain unpopular statements andopinions, criteria for duty of care obligations and their subsequent proceedings must be designed to be particularly sensitive with regard to fundamental rights.

3.3.4. Special Circumstances Close to an Election

Communication regulation is reserved above all for narrow areas of criminally and personality rights relevant statements and depictions with immediate potential for danger. While the existing legal framework already pursues the individual rights-related purposes (see Chapter 3.1.1), the question arises as to legislative possibilities for safeguarding tha knowledge basis for collective decision-making, insofar as this is relevant for the public formation of (political) will and a social negotiation process regarding a statement's truth is not possible for temporal, local or structural reasons. In constellations in which the public discourse with regard to a statement and its truthfulness is not possible, the principle of social self-understanding "tilts" in the direction of a state that compensates for this impossible negotiation process in the event of a concrete threat to legal interests by balancing the freedom of expression and the protection of collective decision-making bases. In these (rare) situations, the state – exceptionally – decides instead of society. Examples of such special arrangements relate above all to statements made in the immediate vicinity of an election that are doubted and are actually able to influence electoral decisions. It should be considered, however, that political information and communication are characterized as an area in which emotional communication takes place and where exaggerations and polemics are common. Persuasion and (legitimate) manipulation are part of the political discourse and "a necessary part of a free election".[212] The threshold of an election-related influence is correspondingly high:

> "An inadmissible influence on the election exists if, in the run-up to the election, public authorities have had a more than insignificant party-biased effect on the formation of the will of the voters, if private third parties, including parties and individual candidates, have influenced the election decision by means of coercion or pressure, or if the formation of the electorate has been influenced in a similarly serious way, without there being a sufficient possibility of defense – for example with the help of the courts or the police – or of compensation, for example with the means of electoral competition."[213]

In particular, the proof that election decisions have been changed on the basis of individual statements, which has not yet been provided (see Chapter 2.2.2), sets constitutional limits to the legislative possibilities here: Only because of the theoretical potential for manipulation, restrictive interventions such as a ban on political statements shortly before an election do not appear justified. It therefore requires an increased potential for manipulation, especially in the form of coercion or (emotional) pressure. Correspondingly strong effects can emanate from statements that are personalized and tie in with individual concerns, fears or political attitudes. This addresses the area of political micro-targeting, in which political messages are disseminated to narrowly defined population groups, and in which the originator hopes that the statement will be perceived in particular.

210 *Lent*, 2020, 593 (599)
211 See in detail *Schierbaum*, [fn. 170], pp. 355.
212 *Klein*, [fn. 55], m.a. 119-120.
213 BVerfGE 103, 111 (132).

3.4. INTERIM CONCLUSION: AREAS OF POTENTIAL COUNTER-MEASURES

The overview of the gaps in protection found in the existing legal framework, as far as the containment of the risk potentials of disinformation is concerned, and the possibilities and limits of legislative action shows that classic regulatory approaches (prohibitions, supervision, sanctions) only seem legitimate for statements that clearly violate individual or group rights. Otherwise such measures are out of the question, especially with regard to statements the truth or falsity of which is not amenable to proof. As a consequence, six major areas arise in which governance measures appear to be legally possible, appropriate and helpful from the point of view of the protection of legal interests. These countermeasures will be analysed more thouroughly in Chapter 4.

3.4.1. Improvement of Regulatory Knowledge

Law, understood as a form of coping with problems by regulating social processes, is not detached from the regulatory area on which it is intended to act. It "standardizes behavioral expectations and creates structural prerequisites for the possibility of norm-compliant behavior".[214] But law can only regulate under the precondition of knowledge. Without knowledge of the current situation, without knowledge of the further course of the development of current situations and in the absence of being able to estimate the consequences of a legal intervention, the development of regulatory measures is not possible. Legislative action presupposes knowledge. As explained in the previous chapters, questions of the dissemination and impact of disinformation on individual and social processes are areas in which relevant knowledge often is not (yet) available. Where this knowledge is missing, one of the first activities of the legislator must be to make a knowledge base available and to process it for own decision-making. Especially in complex regulatory areas with decentralized knowledge sources and complex interrelationships of observable phenomena and (fundamental rights-relevant) effects, the first group of appropriate measures to be taken are those aimed at improving regulatory knowledge (see Chapter 4.1).

3.4.2. Taking Measures in the Event of Statements with Direct Threat Potentials

Where statements directly violate or endanger individual or collective rights, restrictive state measures are not only possible, but also can be deemed necessary. Clearly unlawful and demonstrably untrue allegations can be followed by clear legal countermeasures (see Chapter 4.2); in practice, this applies to extreme cases. In cases of such measures. care must be taken to ensure that freedom of communication is respected and that the requirements of due process are taken into account. The legal determination of the untruth of a statement and the potential for harmful effects requires a case-by-case examination within formal legal procedure. As a consequence, such procedures do hardly scale in view of the quantities and the rate of distribution of potentially damaging expressions, for example on platforms of media intermediaries. However, this is a consequence that applies to all infringing mediated content and cannot be seen as a disinformation-specific aspect.

3.4.3. Enabling and Supporting the Negotiation of Societal Truth Finding

Countermeasures in the event of an ambiguous truth situation, i.e. before an assessment of the untruth in a legal procedure, must be limited to the purpose of enabling and supporting the societal truth finding process - as far as legal measures can be helpful at all here (see Chapter 4.3). Exceptions to the constitutionally indicated reluctance of the state to use the truth/untruth of statements as legal criteria arise where statements are deprived of the process of societal discourse, in particular directly around an electoral act. The same is true for forms of communication that are addressed to a segmented audience but are not accessible to the public, although there is a public interest in the debate about the content of the statement.

214 *Hoffmann-Riem*, 2014, 5.

3.4.4. Counter-factual Stabilization of Trust in Journalistic Content

In special cases, legal requirements can incorporate the adherence of journalistic duties of care with regard to reporting the truth, in order to maintain trust in journalistic content, e.g. in cases in which communicators claim to be truthful, or in which the source of a statement is particularly relied upon because of a supposed claim to accuracy. The respective procedures must necessarily be indeoendent from the state in order to avoid content-related state influence on journalistic content. The degree of adhering to journalistic duties of care depends on the individual case; the yardstick to be applied can be approached based on the criteria of journalistic appearance, the broad impact and the topicality.

3.4.5. Demotivation of Persons Publisihing False Information Primarily Out Of Financial Reasons

Another special case are forms of doubted statements, which are made primarily for economic reasons. In these cases, although the statement is regularly covered by freedom of expression, legal measures aiming at decoupling the statement from the possibilities of its economic use can appear as a proportionate interference with the freedom to exercise one's profession if the statements potentially threaten social legal positions. The decisive aspect here is also the (social) harmfulness of the content. The decision on the impairing effect of an information must be taken before any decision on demonetarization. This leads to a structurally similar problem as with the determination of truth: a uniform case-by-case procedure is required, in the course of which the assessment of falsehood and the necessary balancing can take place.

3.4.6. Prevention of Non-Communicatively Justified Increases of Relevance and Reach

Restrictive legal provisions that tie in with permissible content must be kept to a minimum (see Chapter 3.3). At the same time, a special aspect of disinformation campaigns is also the use (or exploitation) of technical dissemination possibilities, for example by creating a massive number of fake accounts or bot networks that suggest high relevance or strong support for a certain statement. A content-agnostic starting point aiming at ensuring a minimum of equal chances to communicate can therefore be seen in dissemination-related governance measures. One area of countermeasures concerns preventive instruments against such forms of abusive exploitation of digital communication possibilities, which are in particular provided for user-generated content by platforms: Where individual forces influence the perception of expressions so strongly that the pereception of other points of view is considerably more difficult, or these are pushed into the background by artificially created reach, state measures with the aim of (re-) establishing equal opportunities can be considered (see Chapter 4.4).

3.4.7. Meta-challenge: Establishing countermeasures within a framework of "hybrid governance" approaches

When it comes to regulatory approaches implementing countermeasures, a general problem in the governance of intermediaries comes to light[215]: in the area of disinformation governance, the liability privilege of providers who offer third-party content applies (cf. § 10 TMG; Art. 14 EC Directive). In addition, the legislators cannot directly influence (legal) content without violating the principle of independency from the state. As a consequence, neither direct disinformation-related requirements for providers of intermediary services nor forms of revearsel of the burden of proof with a structurally comparable consequence are within the range of legitimate legislative instruments. Especialy the reversal of the burden of proof would lead to the platforms having to refute alleged legal violations by their users, which are neither within their sphere of influence nor regularly within their sphere of knowledge. It becomes clear that the question of the extent to which the burden of proof lies with the platforms in such cases is systematically connected to the question of the fundamental responsibility of platforms for the content of its users.

In addition, there are a number of other circumstances that may influence the choice of adequate governance approaches. For example, the country of origin principle applies to the enforcement of national regulations against information society

215 Cf. *Schulz/Dreyer*, 2020.

services established in other EU countries. Enforceability of any national disinformation-related requirements and enforcement of state measures against these providers can only take place in exceptional cases and in compliance with the procedures according to Art. 3 (4) b), (5) EC Directive. This does not systematically exclude legal action against such providers, but makes the enforcement complex and lengthy. In addition, intermediaries possess far-reaching private autonomy: they can basically determine the form and characteristics of their offer themselves within the framework of general laws, and also change them at any time. This also includes prohibitions and restrictions that go beyond existing and future legal provisions as far as forms of disinformation are concerned. However, where platforms (voluntarily) introduce certain procedures for dealing with forms of disinformation, legal requirements could at least legally frame these procedures and, for example, provide for compliance with minimum standards that safeguard fundamental rights.

Whether and which possibilities the legislator has to regulate platforms directly raises a number of fundamental legal questions. The tensions between form of state and platform-own governance forms in terms of fundamental and human rights have not yet been adequately investigated. The following dimensions play a role here:

- To what extent are platforms bound by fundamental and human rights when developing and applying platform rules?[216]
- What fundamental and human rights protection do the platforms enjoy in the development and application of these rules[217] - also against the background that these rules define their product in the narrower sense?[218] To what extent does the principle of independency from the state apply to the regulation of platforms and communication on platforms?
- To what extent can the state, which sets rules for the platforms, itself be held responsible for subsequent deletions or restrictions on the dissemination of content as a violation of users' fundamental and human rights?

Here, the state moves in a tight field of tension, also with regard to procedural regulations such as the reversal of the burden of proof. For example, the question of the extent to which the burden of proof lies with the platforms can only be decided once the responsibility of platforms for third-party content has been conclusively resolved. Restrictive legal provisions that oblige platforms to deal with disinformation in a certain way can be limited to those forms that result in a threat to legally protected rights. The more tangible this threat is, the more justified legal interventions appear to be. In view of the constant differentiation of the "house rules"[219], it is possible to develop new instruments of state action within the framework of fundamental and human rights - for disinformation and beyond.

Overall, this results in a governance structure for large parts of the regulation of disinformation, in which the state and platforms work closely and cooperatively with each other at best, so that an inner coherence of governance goals and the implementation of the corresponding instrument results. In view of the high level of acceptance and willingness to cooperate required for the measures to be implemented on the part of the platforms, the different factual access options and the necessary scalability of solutions with a view to the sheer number of individual communications, these areas of measures appear as new, modern forms of "hybrid governance"[220]. Here, state regulation and providers' own areas of governance are intertwined and interact with each other (for possible designs, cf. Chapter 4). Whereas forms of self-regulation have so far been viewed rather critically in view of their limited effectiveness[221], forms of co-regulation with state intervention in the background back appear problematic, at least where the governance problems just described reappear as a result of official intervention in the event of self-regulatory failure. Here, new approaches are needed, which in particular allow for state-independent supervision and further development of cooperative rules of conduct submitted by providers.

216 *Mayen*, 2018, 1; *Elsaß et al*, 2017, 234 (238-239); *Engeler*, 2019.
217 Cf. *Gostomzyk*, 2018, pp. 118.
218 *Kettemann/Schulz*, 2020, 33.
219 *Katzenbach*, 2020.
220 „Hybrid forms of governance" are understood as „institutional arrangements" (*Prittwitz*, 2000) that couple hierarchical and cooperative steering instruments, political and technical arenas and thus conflict-laden and consensual policy processes (*Hey et al*., 2008).
221 *Colliver*, 2020.

4. POTENTIAL REGULATORY APPROACHES AND INSTRUMENTS

The identification of structural gaps in legal protection and the analysis of the state's possibilities and limits for closing them has shown that measures in six areas in particular need to be looked at more closely: Those that improve governance knowledge in order to enable legislators and authorities to implement forms of evidence-based governance (Chapter 4.1) as well as those that address individual statements. At this level, different measures can be linked to statements that are objectively and demonstrably false (Chapter 4.2) and those that are not or not completely falsifiable (Chapters 4.3 and 4.3.1), which include the special cases in which a statement contains an increased claim to truth (Chapter 4.3.2) and which are predominantly economically motivated (Chapter 4.3.3). Another potential starting point is the technical form of dissemination of disinformation (Chapter 4.4), see Fig. 4.

Fig. 4: Types ofexpressions and Focus Areas for Countermeasures

Measures to improve regulatory knowledge	• Knowledge on the sides of platforms • Knowledge about effects
Measures in cases of objectively falsifiable statements	• Immitent danger for life and limb, or societal interests or • Misleading statements in direct temporal connection with an election
Measures in cases of doubts about not or not completely falsifiable statements	• Doubts about the truth • Special claim to truth / journalistic appearance • Economically motivated trolling
Measures targeting technological- and distribution-related aspects	• Coordinated inauthentic behaviour • Political communication using microtargeting

Source: Own source.

The analysis of possible governance approaches is carried out systematically and first describes the respective goal (protection of legally protected rights) as well as the measure for achieving this goal. It is then analyzed on an actor-by-actor basis which addressees come into consideration when it comes to the implementation of the measure, which regulatory approach seems promising and to what extent the success of the measure depends on certain (pre-) conditions when it comes to effectiveness.

4.1. MEASURES TO IMPROVE REGULATORY KNOWLEDGE

In order to be able to react effectively and appropriately to risk potentials of disinformation, state players and social actors need knowledge. The required knowledge basis refers to knowledge about the extent and dissemination of relevant statements and campaigns, knowledge about the involved player structures and actor networks, empirical evidence on the effects and potential effects of statements in question, but also knowledge about the forms, scope and effects of implemented tools and measures of private players in this area.[222] If the state decides on legal or regulatory measures on the basis of the status quo, the knowledge about the effects and consequences of any intervention is also relevant. With regard to these very different sets of knowledge, it becomes clear that the relevant knowledge is not available centrally but can only be gained de-centrally from different players, or must be produced there. As measures for the acquisition of regulatory knowledge, forms of disclosure, transparency obligations or access rights to existing bodies of knowledge on the side of private entities can be considered.

With regard to knowledge about the scope and reach of disinformation phenomena, about actor structures, (dis)information content and communication strategies as well as about implemented countermeasures, providers of media intermediaries are coming into focus. Access to their body of knowledge about public and semi-public communication processes on their platforms (platform knowledge) thus appears to be the central precondition of good governance. Potentiallly obliged parties would be private actors, who may be entitled to invoke trade secrets in the case of disclosure obligations. Another structural disadvantage of information obligations or access rights is that a critical review regarding the accuracy of data sets stemming from the sphere of private companies regularly is not possible. When it comes to ensuring the reliability and validity of the data and information provided by platform providers, however, traditional legal obligations reach their limits. This is an example for the need for close cooperation between the state and private actors in this area, as outlined in chapter 3.4.

Insights into the processes of the platforms provide regulators as well as scientists and the public with the possibility of assessing the extent of disinformation activities and how efficient platforms are with regard to their own platform-related governance decisions concerning individual statements or accounts. By disclosing the number of content moderation cases and decisions, the providers themselves can prevent false expectations and provide the basis for targeted regulatory approaches.[223] As a regulatory measure, for example, an obligation of the platform to publish regular transparency reports for the area of disinformation comes into consideration. Already existing monitoring obligations point to optimisation possibilities with regard to comparable report structures, uniform criteria and counting procedures as well as coherent areas of observation. A further possibility is the explicit obligation to provide country-specific data in order to avoid having to use only aggregated transnational figures as a basis for decision-making. Furthermore, such reports should go beyond pure case numbers and, in particular, show identified larger campaigns and currently applied (dis)information strategies as well as make the platform's internal procedures and decision-making processes transparent in an encompassing manner.

However, the knowledge available on the side of private companies cannot yet make a direct statement about potential risks or effects on the recipients's side; such knowledge is usually generated by science. The scientific production of knowledge regarding effects is an important prerequisite for the assessment of possible state countermeasures, too. This circumstance clearly shows the relevance of researchers' access to the platforms' bodies of knowledge. Where platforms invoke regulatory obstacles such as GDPR requirements when granting access to data, the possibilities of research privileges under data protection law, as set out in Article 89 GDPR, must be fully seized. By concretising the opening clause in individual Member States' statutory laws, further legal clarifications can be made. Further legal bases can also be created at EU level: the draft Digital Services Act contains the obligation of very large online platforms to grant data access for researchers in Article 31 (2). Here, however, regulatory activities are not limited to granting access to scientists, but can also encompass funding activities for research projects and formats of scientific knowledge transfer into policy arenas.

Finally, legal measures must also be discussed in view of their intended effects and their possible consequences for protected legal rights before they are implemented. The necessary knowledge about regulatory impact comes either from the experience of the regulator or is based on regulatory science and evaluation research. As a rule, the legislator already has access to the (limited, because future-oriented) knowledge. At this level, it is primarily a matter of guidelines to keep such knowledge available and to be aware of using it and taking it into account when making policy decisions.

222 Cf. the proposals for a research agenda in *Bliss et al*, 2020.
223 *Gorwa/Ash*, 2020, 295 (incl. further ref.); *Keller/Leersen*, 2020, pp. 220.

4.2. MEASURES IN CASES OF OBJECTIVELY FALSIFIABLE STATEMENTS

From a legal point of view, the truth of a statement can only be used to a very limited extent as a criterion for measures, since truth is primarily constructed by society in (communicative) negotiation processes (see Chapter 3.3). Truth can therefore only form a legal starting point where statements can undoubtedly and objectively falsified, for example in cases of obvious deviations from an objectively observable state, in the case of completely invented statements or provable manipulations of photographic, audiovisual or audio recordings (manipulated content)[224], and where legally protected rights of individuals or society are infringed or endangered as a result. For this purpose, uniform (court) procedures on a case-by-case basis are necessary, the aim of which is to finding and determining the truth and which, on this basis, judge about legal violations, consequences and resprective claims.

4.2.1. Statutory Prohibitions

The legal framework currently prohibits untrue statements within narrow limits, see §§ 185 ff., 130 StGB. Outside of these exceptions, the dissemination of disinformation regularly does not constitute a criminally relevant violation of the law.[225] The reason for this is that the process of negotiating the truth lies predominantly with society. A principal legal prohibition of the utterance of untrue claims is ruled out simply because then a (central) state authority would have to decide on the truth or untruth of a statement. Where possible untrue allegations endanger individual or social legal positions, however, legal measures can be considered[226] – but criminal sanctions can only be applied to conduct that is significantly damaging society.[227] This requirement is met if statements are particularly reprehensible or pose significant risk potential;[228] this is the case, for example, when defamatory false statements are made about individuals or groups (§ 187 StGB). These narrow limits on prohibited statements are also necessary in light of freedom of expression, since general prohibitions in particular can constitute a significant interference with freedom of expression.

With regard to the legal protection gaps identified in Chapter 3, there are three areas in which a legal prohibition of statements appears possible in principle: (a) with regard to disinforming statements, which constitute a direct threat to the life and limb of a third person;[229] (b) with regard to falsifiable claims that pose an immediate, concrete threat to social interests such as public safety or public health;[230] and (c) in relation to misleading statements in direct temporal connection with elections in which, for reasons of time, a social negotiation of the truth is not possible. The latter situation can arise in particular in the case of disinformation, which is spread in the immediate vicinity of an electoral act and which show an increased potential for manipulation (in particular through coercion or emotional pressure), since the absence of appropriate discursive counter-speech could endanger the freedom and integrity of the election.

Such statutory prohibitions directly oblige the person making a statement; however, via the liability requirements for platforms hosting user-generated content (host provider, § 10 TMG), these providers are also responsible for such content from the time of obtaining knowledge of a legal violation. Indirectly, corresponding prohibitions can thus be addressed to platforms in the form of deletion obligations.

4.2.2. Deletion/Blocking

In parallel with the prohibition of such information, its deletion or blocking can comeinto question. By deleting disinforming content, the identified situations with currently lower legal protection can also be improved, since the threat to protected legal rights usually decrease with lower perceptibility and reduced dissemination. A special criterion of the three situations mentioned is that there is imminent danger, i.e. it is crucial to remove the content as fast as possible. In enforcing the prohibition, exceptional forms of enforcement are possible, where the relevant statements are deleted or blocked first, and

224 So-called deep fakes, c.f. fn. 17.
225 Issues arise regarding the enforcement of the existing criminal provisions online, which are supposed to be addressed by the NetzDG; cf. *Spindler et al*, 2018, m.a. 6.
226 *Helm/Nasu*, 2021, 302.
227 Cf. BVerfG, NJW 1998, 443 (443).
228 *Löber/Roßnagel*, [fn. 71], 182.
229 *Helm/Nasu*, 2021, 302.
230 *Helm/Nasu*, ibid.

their falsifiability is only subsequently checked in the context of legal proceedings (concerning further aspects of limiting visibility see the paragraph on donwranking/depriorisation).

With regard to freedom of expression, the measure of content deletion should in principle be regarded as a restrictive instrument. A legal obligation to delete content may be indicated in particular where statements violate a legal prohibition[231] - here, the deletion or blocking can be seen as an accessory measure in enforcing a statutory prohibition. On the other hand, a state-ordered obligation to delete all objectively falsifiable content is illicit, since a state decision on truth can have potentially serious effects on freedom of expression (see Chapter 3.3).

However, platforms themselves could reserve the right to delete content in their terms and conditions that is objectively falsifiable, even without showing risk potentials.[232] In the context of the application of provisions in the terms and conditions by the platform, the horizontal effect of freedom of expression pursuant to Article 5 GG as well as the general right to personality of users pursuant to Art. 2 Sect. 1 in conjunction with Art. 1 sect. 1 GG on the general clauses of §§ 242, 307 BGB has to be taken into account. However, under German law, false factual claims in particular are excluded from the scope of protection of freedom of expression.[233] In addition, the deletion of *objectively falsifiable* statements represents only a minor interference with the freedom of expression of the affected users. The deletion procedures that platforms apply in such cases would have to be uniform regarding the consideration of the horizontal effect of the fundamental rights, so that an arbitrary deletion of content is ruled out. In addition, comlaint or appeal proceedings should be provided for persons affected, giving them the opportunity to have a deletion or blocking decisions reviewed and, if necessary, to implement its annulment. This is particularly necessary in view of content that has been incorrectly identified as untrue, since such wrong decisions can be seen as relevant limitations of freedom of expression.

Easily accessible and comprehensible information on the reasons for removal and the possibilities for appeal provide a basis for reviewing and contesting the decisions.[234] In addition, these forms of disclosures and reasonings can also lead to users being more likely to accept or understand the decision and not to create or distribute disinforming content in the future. If the decisions on the basis of such transparency measures can be verified by the users or, if necessary, external third parties, trust in the platform could also increase. Accordingly, regular review mechanisms could also be established at this level.[235]

4.2.3. Downranking/Deprioritisation

The measure of deprioritisation of content hooks into the attention economy-driven logic of the platform: In the context of selection and prioritisation procedures, which filter and sort content to be displayed, certain content can be deliberately made less visible. Disinforming content, which is displayed less in the feeds of the users, is therefore less perceptible, less received and less disseminated.[236] The legal threats posed by information would be significantly reduced by deprioritisation. Since deprioritised content is in fact deprived of its visibility in the discourse and thus structurally resembles deletion in its consequence with regard to the individual communication, a legal obligation to do so only comes into consideration in the case of objectively falsifiable content with risk potential. Against the background that a statutory obligation to delete untrue information is not eligible (see above), this must also apply to deprioritisation due to its factually comparable effect.

However, platforms could reserve the right in their terms and conditions to deprioritise objectively falsifiable content, independent of their risks for legally protected rights. The selection logics are designed to show or recommend to the users those content that could be of particular interest to them. Any prioritization of one piece of content means the deprioritisation other content anyway. If a provider reserves the right to deprioritise falsified content, this is possible within the framework of contractual private autonomy (but it is limited by § 94 MStV in caes of deprioritisation of journalistic content). When

231 See § 3 II Nr. 2 NetzDG.
232 So far, only a few platforms exclude any form of disinformation in their terms of use, and even then only if it has the potential to cause harm (Pinterest, TikTok). Most providers, including Facebook, Twitter and YouTube, reserve the right to sanction disinformation only in certain cases and contexts, such as potential violent consequences or influence on elections. Cf. *Bateman et al*, 2021.
233 Grabenwarter, [fn. 55] m.a. 49.
234 *Jankowicz/Pierson*, 2020, 9.
235 *Ash et al*, 2018, 1 (19).
236 See *Caplan et al*, 2018, 21.

applying appropriate procedures, the platform must take into account the horizontal effect of human rights; arbitrary or objectively unjustified deprioritisations are inadmissible in this respect. In addition, the platform itself is responsible for determining the falsifiability of a statement; this is problematic from the point of view of the principle that negotiating the truth is principally handed over to social discourse.

In addition, deprioritisations can be more difficult to detect; those affected may be deprived of effectively making use of complaint mechanisms.[237] An indication that deprioritisation has taken place could be labeling the relevant content as disputed or untrue. With a view to the design of the deprioritisation procedure in order to safeguard fundamental rights, an obligation could thus arise to label the content in addition to deprioritisation in order to enable the affected users to object to the deprioritisation decision. However, it seems preferable that the platform informs users about a disinformation-related deprioritisation directly. In addition, the procedural requirements discussed in the context of deletion also are indicated here. Complaint mechanisms should be implemented and easily accessible while comprehensible information on any platform decision in this context should be provided.

4.2.4. Labels

There are conceivable constellations of cases in which transparency in the sense of making objectively falsifiable content recognisable may appear more expedient than its removal. This is particularly the case where forms of labelling use potentials of awareness-building, especially in cases where the content in question or the person stating it holds a special position in the social discourse.[238] If objectively falsified content is perceived as such by respective labels, the untruth can be corrected accordingly when it comes to the world views of the recipients. Moreover, labelled statements are also still available for social and individual engagement with the content.[239] In contrast to "only" doubted or disputed assertions (see Chapter 4.3.1.2), it seems possible for cases case of content that has been objectively falsified to make clear the (proven) untruth of a statement that otherwise does not violate any legal rights, for example by labelling it as "untrue" or "refuted". The requirements the process must meet when applying such labels are shown in Chapter 4.3.1.2.

4.3. MEASURES IN CASES OF DOUBTS ABOUT NOT OR NOT COMPLETELY FALSIFIABLE STATEMENTS

Much of the disinforming content is not - or not completely - falsifiable. As a rule, they cannot be countered with restrictive requirements such as prohibitions or deletions, but the negotiation of their true and untrue contents is primarily handed over to social discourse (see Chapter 3.3.1). The possibilities for regulatory action are therefore limited to enabling and promoting such social negotiation processes regarding the truth of expressions; these negotiation processes take place through discourse. Measures to strengthen discourse, i.e. those that enable or promote discourse, can therefore have a positive effect on the social process of finding truth.

4.3.1. Measures in cases of challenged statements

The core of the deliberation process within the social discourse is the interaction of speech and counter-speech. The perception of divergent positions and the confrontation with other viewpoints are central components. This interrelationship is promoted when the counter-speech is made visible and leads to an engagement with it. The social negotiation process is strengthened if doubts about their truthfulness are made visible in the case of content that cannot be objectively falsified, and thus the claim to truth of the statement is disputed.[240]

237 *Douek,* 2019, 1 (18)
238 Discussed with regard to the deplatforming of Donald Trump: *Reinhardt*, 2021; *Fertmann/Potthast*, https://www.juwiss.de/05-2021/.
239 This refers to the fundamental problem of profile deletions, in which relevant content is removed as documentation material for the discourse.
240 „People typically assume what others say is truthful and accurate unless there is reason for doubt", *Britt et al*, 94 (98) (incl. further ref.).

4.3.1.1 Identification of Dubious Statements

In order to make doubts visible, the first step is to identify potentially untrue or dubious content. These identification measures accordingly can be seen as the starting point for a number of other measures, which are presented below.

Reporting Options

First, identification of dubious content can be made enabled by reporting mechanisms, which are also referred to as "flagging". Reporting mechanisms enable users or other actors to report content;[241] the report can then be followed by a (preferably external) procedure to check the truth of the information or a form of labelling, for example in cases of frequently reported content (cf. following sections).

Reporting mechanisms can be designed differently, especially with regard to the persons and actors submitting reports. First, users of the respective platform can report potential disinformation. Such user flagging functions is currently common practice on many social media platforms. Facebook and Instagram, for example, allow users to report a content as a possible violation of the terms of use[242] on the grounds that it is false information.[243]

Concerns with regard to user flagging arise from the fact that users may not have sufficient interest or expertise to reliably report untrue content.[244] In addition, it is noted that user flagging systems can be exploited (in coordinated ways) to systematically report certain persons or opinions and thus, in the worst case, exclude them from the discourse.[245] However, this risk can and is often mitigated by the fact that further measures are taken only after a cursory examination of the truthfulness of the respective information (see Chapter 4.3.1.2).

A special form of flagging is the reporting of content by experts or particularly trustworthy institutions or users[246] (trusted flagging). YouTube, Facebook and Twitter have implemented such trusted flagger programmes, in which non-governmental organizations, government agencies and individual users, who ideally have expertise in at least one of the areas prohibited by the Community Guidelines, can report content. Messages by trusted flaggers are then checked in a prioritised manner by the platform (see Chapter 4.3.1.2).[247] This form of reporting by experts and particularly trustworthy institutions and users offers the advantage that due to the expertise and/or the reliability of the actors, there is a lower risk of incorrectly reported statements. Potentially, this can also save resources during subsequent checks. However, on YouTube so far less than 1% of the content is reported by trusted flaggers[248], which illustrates the limited scalability of expert-based flagging programs. Nevertheless, trusted flagger programmes and initiatives are in principle suitable for receiving indications of relevant false information from expert actors, while the further handling of the reports is left to the platforms.

Reporting functions for individual users are already provided by most platforms, too. A legal obligation to maintain and implement reporting functions can in principle also be laid down by law. Reporting functions, which enable individual users or experts to report content to the platform provider (not to label it) do not yet constitute an interference with the freedom of expression or the personal rights of the affected users. The associated intervention into the right of freedom of profession is regularly justified in view of the risk potentials arising from disinformation.[249] Where platforms already provide procedures for dealing with user reports, legislation may provide rules on minimum standards to be met, including complaint mechanisms.

In order to ensure the effectiveness of flagging functionalities, certain process designs are suitable. Notifications by users require their knowledge about the existence and procedures of such possibilities. The platform can inform about this, for example, by means of information at the time of the user registration or by means of easily understandable symbols and icons during user sessions. Since the reports by users or experts are usually followed by a content review by the platform,

241 The option to report content is widely available on social media platforms; Facebook: https://de-de.facebook.com/help/1380418588640631; Twitter: https://help.twitter.com/de/safety-and-security/report-abusive-behavior; Instagram: https://about.instagram.com/de-de/community/safety.
242 The distribution of false information is one of several reasons for reporting.
243 The terms used by the platforms vary here between "false information" or "misinformation".
244 See in context to the reporting of hate speech: *Wilson/Land*, 2020, 1 (36)
245 *Wilson/Land*, 2020, 1 (36).
246 "Users who reliably report a large number of videos can participate in the trusted flagger program", https://support.google.com/youtube/answer/7554338.
247 Youtube Trusted Flagger Program: https://support.google.com/youtube/answer/7554338.
248 YouTube Transparency Report, https://transparencyreport.google.com/youtube-policy/flags
249 See existing applicable provision § 5a Sect. 2 P. 2 JMStV.

it can also be helpful in terms of quality and scalability to hire a larger number of contextually savvy (e.g. with regard to Covid-19 in the healthcare sector) content reviewers.[250]

Types of Automated Content Recognition

Finally, there is also the possibility of proactively detecting disinformation by the platform itself. First and foremost, this is currently done through the use of algorithms that pre-filter possible relevant content in order to have it checked by humans.[251] A distinction must be made here between algorithms that filter out textual content and those that are used to uncover manipulated images and audiovisual material. Textual content can be examined automatically either through approaches that use text analysis methods or through the use of metadata.[252] In the case of images or audiovisual content, on the other hand, a distinction is made between matching models and predictive models. The former compare content with a comprehensive database and identify similarities with existing data. The latter are used to identify features of new content that has not been assessed before.[253]

Automated assessments of the veracity of content (for example, based on its source or URL) or differentiation according to its reliability (based on an evaluation of its creators; see Chapter 4.3.1.2.)[254] are demanding; but at least they ignore the doubts regarding the suitability of "truth" as a starting point (see Chapter 2.1.2). Hoewever, it is questionable to what extent automated procedures can consider the context necessary for a nuanced assessment of content into their decision. It is to be feared that any automated reviews will sometimes identify as disinformation those contents that have been expressed in a legitimate manner ("false positives"). As a rule, such measures constitute an unjustified interference with freedom of expression pursuant to Article 5 sect. 1 p. 1 GG. The real aim of these measures would therefore be thwarted.

Even with the automated detection of image or video manipulations, false positives can result. This suggests that algorithms can only make a pre-selection for a human review - comparable to user reports - and cannot independently filter out disinformation.[255]

Time of Automated Recognition

An early identification of disinformation is crucial to prevent further dissemination and thus a possible far-reaching influence through disinformation (for the dissemination of disinformation s. Chapter 2.1.1).[256]

An early detection, for example through so-called upload filters, is often difficult: a legal obligation to use upload filters is not legitimate as it would contradict human rights as well as the prohibition of general monitoring obligations in Art. 14 EC Directive. The possibility of using voluntary preventive approaches is usually not taken up by host providers, as this would also mean that the time of knowledge and thus any potential liability would occur earlier in the case of platforms with user-generated content (§ 10 TMG). In principle, however, platforms could implement such preventive measures voluntarily on the basis of their private autonomy. In the future, a provision such as the draft Art. 6 DSA may help in minimising liability risks for voluntarily implemented detection procedures. With regard to the wording and the procedures laid down in the terms and conditions, the horizontal effect of fundamental rights via §§ 242, 307 BGB have to be taken into consideration. In particular, such procedures should be uniform and transparent and provide a complaint procedure. This is because the platform receives the sovereignty of interpretation over what is classified as "disinforming content" when implementing upload filters. This could also unduly filter out legitimate expressions. A procedure that safeguards fundamental rights depends, inter alia, on the quality of the automated recognition of doubtful statements.

4.3.1.2 Verification Processes and Indication of Doubts

Once relevant content has been identified, a number of other measures can tie in with this circumstance. As a direct follow-up measure, fact checking comes into consideration, in which content is checked for its truthfulness.[257] Such a review can be a further (possibly optional) intermediate step before measures with the aim of *making doubts visible* are used.

250 In connection with hate speech in particular with regards to special cultural and political situations: *Ash et al*, [fn. 219], 1 (19).
251 *Camebridge Consultants*, 2019, 6.
252 *Halvani et al*, 2020, 102.
253 *Shenkman et al*, 2021, pp. 12.
254 See *Kyza et al*, 2020, 1 (16).
255 *Halvani et al*, [fn. 236], 141-142.
256 *Shu et al*, 2020, (5); Brashier et al, 2021, 1.
257 See *Walter et al*, 2020, 1.

These measures may consist in particular in tagging or (linking of) refuting information. By identifying doubts, a strengthening of discourse can be effected, which can counteract the risk potentials raised by disinformation for the individual and collective (political) formation of opinion as well as threats to the integrity of elections. The aim of the identification of doubts is to make users aware of the possibly untrue content of a statement and thus ideally to cause a critical (re-) assessment of an expression by the recipients. As long as the relevant statements are lawful and do not violate the rights of third parties, a legally binding general requirement for platforms to maintain content review procedures seems – also with regard to the liability privilege of Art. 14 EC Directive – disproportionate; however, the overlap of disinformation and hate speech can lead to factual overlaps with legal obligations to provide reporting mechanisms with regard to unlawful content (in particular § 3 NetzDG, § 10a TMG).

Due to their private autonomy, platforms are entitled to stipulate such measures in their terms and conditions. The legislator is free to legally oblige platforms to comply with certain procedural requirements for the corresponding measures if this option is used.[258] A legal provision regarding any procedural requirements to be fulfilled can offer the advantage of creating a cross-platform minimum standard here, even if the concrete implementation may vary from platform to platform. However, corresponding legal regulations must be within the limits of proportionality and contain exceptions, in particular with regard to feasability, e.g. with regard to very small platforms. Fact checking is associated with a high level of personnel and financial expenditure, which can quickly threaten the existence of very small or non-profit platforms, which is why a corresponding legal obligation to provide appropriate procedures might slip into the area of disproportionality.

If no regulations are provided for by law, the horizontal effect of the fundamental rights must be taken into account with regard to the procedural design and application of the respective measure. Although the fundamental rights do not bind private actors directly, they have an indirect effect via the general clauses under civil law (such as § 138, 242 and 307 BGB), which oblige them to interpret these general clauses in accordance with fundamental rights.[259]

It follows that if certain measures are stipulated in the terms and conditions of platforms, requirements with regard to their procedures must be met and these must in particular be uniformly designed. In this context, the question must be asked as to whether the respective terms and conditions or their application in the specific case establish a balance between the fundamental rights concerned. This may be the case in general or due to the concrete design, specifically, if terms and conditions are too vague, or if their application is arbitrary, discriminatory or, due to their design, unduly restricts the communicative conditions as a whole.[260]

Fact Checking

Fact checking describes a procedure in which content is systematically checked for its truthfulness in order to determine whether it corresponds to the facts or if it can be refuted.[261] Usually the result of such a review is published.[262] Fact checking may be conducted by different players, in particular the hosting platform itself[263], independent civil society institutions or established media companies.[264] Institutions and initiatives in Germany are, for example, dpa-Faktencheck, CORRECTIV, ARD Faktenfinder, BR Faktenfuchs, or HOAXmap. Internationally[265], initiatives such as FactCheck.org, AFP Fact Check, Snopes, Mimikama and those of established media such as the Washington Post Fact Checker are well known.

In the context of platforms, fact checking is regularly only the first step towards measures such as tagging and labelling. For these subsequent measures, an overall review result can therefore only be doubtful or not doubtful. However, it follows from this that statements containing several facts are often (and must be) handled as if they contained only one fact.[266] This means that, for example, political statements that may contain a variety of facts (and their interpretations) can only be treated as true or untrue.[267] As a result, unquestionable parts of a (political) statement may also be found to be

258 See Art. 12 Digital Services Act (proposal).
259 Armbrüster, 2018, m.a. 34 (incl. further ref.).
260 Kühling, 2018, m.a. 38b (incl. further ref.).
261 Based on the definition of *Walter et al*, [fn. 241], 1 (2).: „Fact checking is the practice of systematically publishing assessments of the validity of claims made by public officials and institutions with an explicit attempt to identify whether a claim is factual."
262 *Walter et al*, ibid., 1 (2).
263 In reference to Facebook cf. *Lazer et al*, 2018, 1094 (1096).
264 *Lazer et al*, ibid., 1094 (1096).
265 There is a comprehensive overview of fact checking pages on https://en.wikipedia.org/wiki/List_of_fact-checking_websites.
266 *Uscinski/Butler*, 2013, 162 (163-164).
267 *Uscinski/Butler*, 2013, 162 (163-164).

labelled doubtful, which in turn could have a negative effect on the knowledge basis and the formation of opinion. This risk could be counteracted, for example, by indicating that only parts of a statement may be doubtful or disputed - or refuting or official information is linked, that can make clear which part of a statement is dubious.[268]

In addition, concerns are raised about the cost and scalability of fact checking.[269] Fact checking is associated with significant costs and so far only a small proportion of new content can be checked for its correctness.[270] Here, cost advantages and improvements in effectiveness can be achieved through cross-platform, interoperable approaches.[271]

If platforms implement such fact-checking procedures (including by commissioning independent organizations) in their processes, these procedures must be designed in a way that safeguards fundamental rights. Precisely because of the sensitivity to fundamental rights of checking the truthfulness of individual statements with regard to the freedom of expression (Art. 5 sect. 1 p. 1 GG) and the requirement of independency from the state,[272] fact-checking procedures must meet certain requirements with regard to their procedure, internal organisation, stakeholder participation and financing in order to be designed to safeguard fundamental rights.

The requirement to act independent from the state (see Chapter 3.3.1) first of all requires that the state and its actors must not have a "decisive influence" on media content.[273] First of all, it follows that fact checking must be carried out by non-governmental and independent organizations. When the independence of such an organization can be assumed is measured by several components, such as the actual entity, the legal framework and the financing. Based on the requirement of independency from the state, all three components must be designed in such a way that no determining influence can be exercised by the state or individual actors with regard to content. In concrete terms, it follows that, in particular, integration into state-owned bodies, direct state financing and an overly narrow legal framework cannot meet these requirements.

In addition, fact-checking procedures and their consequences (especially in form of subsequent measures) must be designef in a transparent way in order to ensure their verifiability. This also includes their traceability and thus strengthens the trust of affected persons, users and the public. This process should be easily accessible and explained in simple language. Furthermore, the establishment of an external complaints body may be useful. This also ensures independent review and decision-making procedures.

In addition, forums consisting of several relevant stakeholders[274] such as representatives of platforms and independent fact-checking organizations as well as decision-makers can also be a suitable instrument to make fact checking more effective. The same applies to cross-platform communication and exchange forums. Through the exchange of experience and expertise, these can contribute to the development and optimization of fact-checking processes and, if necessary, enable cross-platform forms of cooperation and interoperability of fact checking results.

Labelling of Questioned Content (Tagging)

One measure is the labelling of dubious or disputed statements and information. Such labelling is also known as tagging (sometimes also called flagging). The term "tag" comes from computer science and describes the labelling of a content with further information.[275] In the following, the term "tagging" is used.

Tagging measures are already part of the disinformation strategies of some platforms and often embedded in a tiered system of measures.[276] Tags can be designed differently in social networks, on the one hand with regard to the actors who raise doubts, and on the other hand with regard to the design of the label. The aim of such signs is to draw the user's attention to the possibly untrue content of a statement and thus ideally to cause a critical assessment of the content by the recipients. The labeling of a content can be preceded by various measures (e.g. assessment by a platform or by fact-checking organisations).

268 Corresponding practices already exist on individual platforms, but only temporarily, e.g. during election campaigns.
269 Cf. the analysis of the Stiftung Neue Verantwortung, *Sängerlaub*, 2018.
270 *Caplan et al*, 2018, 18.
271 *Sängerlaub*, 2018, 20.
272 BVerfGE 121, 30 (50, 53).
273 BVerfGE 121, 30 (50, 53).
274 Corresponding multi-stakeholder forums are also deployed under the EU Code of Practice ("Multistakeholder Forum on Disinformation").
275 See https://www.duden.de/rechtschreibung/Tag_Strukturelement_Markierung.
276 For instance, in March 2020, Twitter changed its policy to flag misleading or controversial content on Covid-19. See https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information.html.

In addition, tags can be created by users, especially by forms of "crowdsourcing".[277] Here, community-based mechanisms can be used[278], where tags are displayed from a certain number of user reports. However, complicity of users or the use of (social) bots in particular can result in considerable risks of abuse. These risks could be mitigated by diversity requirements regarding the consensus expressed by user notifications;[279] however, difficulties arise here in determining both the notion of diversity and determining a threshold. It also seems possible to apply labels on basis of taggings made by experts or trusted flaggers. However, due to their limited number there is limited scalability (see Chapter 4.3.1.1).

In addition, different design options are possible. Content can, for example, be marked by tags as controversial;[280] the tags can be designed as a (pre-)warning or reference to a contradiction to official information[281] or merely contain references to public information.[282]

In addition, the context of information can also form a criterion for determining whether and to what extent content is tagged. Twitter, for example, takes special action against disinformation related to Covid-19 by either deleting it, applying a warning to a statement or by labeling it, based on the probability of harmful effects.[283] In this context, it is also possible to extend all statements that contain certain keywords by including links to public information.[284] Potentially, however, this could have a lower corrective effect on the recipients, since such tags might be displayed on the one hand very frequently and on the other hand independently of existing doubts as to the truthfulness of a specific statement.

Another special case of making doubts visible are (pre-)warnings. Here, a warning is placed before users perceive the respective content.[285] This measure can therefore potentially cause a critical attitude of the recipients even before a content is perceived and, in some cases, even prevent a dubious content from being perceived at all. Such measures can be implemented in the context of platforms, for example, by applying warning banners in cases of sensitive content. Here, the content for users is obscured by a banner which must be clicked before the content becomes visible.[286]

If labelling measures are stipulated by the platforms in their terms and conditions, their procedures must meet certain requirements and be uniform. As a procedural requirement for labelling it should in particular be ensured that a complaint procedure is implemented with which affected users can take action against any incorrectly made labelling decisions. In addition, the procedure by which statements are being tagged must disclose the reasons for the individual decision, as well as all necessary information on the complaints procedure, in clear and simple language. This is the only way to make it possible to review the respective decision and to create trust in the platform. Similarly, the legislator might also define such procedural requirements for the platform in a binding manner.

Addition of Confuting Information

Another (complementary) way to achieve a critical assessment of false information is to include confuting or disproving information. This debunking also aims at making users aware of the dubiousness of a statement, including possible counter evidence, thus causing a critical assessment on the part of the recipients.

Refutation can have different appearances: On the one hand, these can be attached to the disputed information by a simple link.[287] Such refutations about widespread disinformation are usually provided by fact-checking organizations,[288] which means that this measure is regularly linked to a previous fact check.

277 „Crowdsourced flags"; see *Gaozhao*, 2021, 1 (3)
278 As an example Twitter's *Birdwatch* can be mentioned, https://blog.twitter.com/en_us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation.html.
279 See Twitter's *Birdwatch*, https://blog.twitter.com/en_us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation.html.
280 *Rini*, 2017, E-57.
281 See Twitter, Updating our approach to misleading information, 11.05.2020, https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information.html
282 Cf. the practice of Instagram, https://help.instagram.com/975917226081685
283 See Twitter, Updating our approach to misleading information, 11.05.2020, https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information.html
284 In the current pandemic situation, it is common practice of Instagram to apply tags if the word "vaccination" is used.
285 So-called „prebunkings"; e.g. *Brashier et al*, [fn. 240], 1 (1).
286 This is for example common practice at Instagram; https://www.facebook.com/help/instagram/188848648282410.
287 For instance proposed by *Rini*, 2017, E-56.
288 *Hameleers*, 2020, 1 (1-2).

In addition, refutation could also be provided by particularly verified users (similar to trusted flaggers). In this context, however, it has been suggested that these reports should also contain links to information about fact-checking and to information corrected by official authorities.[289]

In addition, crowdsourcing approaches can also be used. Twitter, for example, is testing a community-based system for contextualizing disinformation. This is intended to enable users to identify information in tweets that they consider misleading and to add comments on the context. If a consensus emerges from "a broad and diverse group of contributors", comments should finally be made globally visible directly in the tweets.[290] However, this raises the question of the number of notifications from which a "consensus" can be assumed and how the diversity within this group can be determined. Through malicious cooperation, there is also a risk of abuse here, possibly imapairing undesirable opinions and perspectives. With regard to such forms of cooperations and the use of social bots, misuse potentials also exists when setting comparatively high thresholds.

As a procedural requirement, it should also be ensured that the procedure of adding a refutation is transparent for the users and that a complaint procedure objecting a refutation is being provided. As part of this, the affected users must be enabled to present evidence against the confuting information and thereby, if necessary, to achieve the removal of it. One advantage that refutation offers is that this already makes visible what doubts exist specifically and that differentiated contributions and evidences can be linked.

Labelling on the Level of User Accounts

Also at the account or profile level, labels and tags can be used as a countermeasure, e.g., if dubious information has been shared several times via certain accounts. Accordingly, a so-called reputation score has been proposed as a measure. A reputation score is the calculation of an individual value "based on the frequency with which each user has shared controversial stories".[291] A label of relevant users could be based on colored icons, including the reputation value, next to user photos[292] or other transparency-directed tools, if necessary without specifying the exact value.[293] The aim of such a measure is to give the recipients an impression of the credibility of the person or source making a statement and thereby to enable a critical assessment of the respective content.

The labelling of user accounts in the event of multiple cases of disseminating dubious information – provided that the platform stipulates it in the the terms of services and not regulated by law – seems permissible. With regard to the arrangement of such a measure, especially regarding the procedure, the horizontal effect of fundamental rights must be taken intoaccount here, too. Accordingly, the threshold for tagging a whole profile must not be set too low, in particular because we aretalkgin about not objectively falsifiable statements. In addition, the dissemination of dubious information can only be used as an indicator for creators or distributors – and not for accounts simply sharing or forwarding such a statement. If a content is shared or forwarded by an account, it cannot be assumed that the information was disseminated with the knowledge of its dubious truthfulness. In addition, depending on the design of the measure, there is a potential for abuse. For example, if a reputation score is based on community-based reporting mechanisms, this can be exploited to unfairly impair certain points of view and people. This potential for abuse can be countered in particular by the fact that the process of determining the truthfulness is carried out by independent actors beforehand. If, for example, a label is applied in the case of ten reports concerning "controversial" content of an account, it is necessary that any account-wide label is based on an independent fact checking (with regard to the independency of such bodies see the paragraph on factchecking above).

As further procedural requirements, it should be ensured that users have access to a justification for the labeling and that there is a possibility of complaint against the labelling decision. In addition, clear information insimple languge regarding the labelling procedure as well as regarding the complaint options must be established.

289 See *Kyza et al*, [fn. 238], 1 (15).
290 Regarding the *Birdwatch* Initiative cf. https://blog.twitter.com/en_us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation.html.
291 *Rini*, 2017, E 57.
292 See *Rini*, 2017, E 56.
293 Facebook announced a similar course of action in May 2021, https://about.fb.com/news/2021/05/taking- action-against-people-who-repeatedly-share-misinformation/

Effects of (Pre-)Warnings, Labels and Confuting Information

The effects of (pre-)warnings, labels and/or refuting information have not yet been sufficiently researched. Initial findings show a range of possible effects, ranging from moderately positive results to backfire effects, which counteracts the actually hoped-for effects of such measures.[294] The research build on findings about human behavior in dealing with information and memory functions.

People generally tend to prefer and accept information as more credible and more convincing if these are coherent with their existing beliefs (confirmation bias).[295] They are also more likely to look for information that confirms existing opinions[296] and remember such content better[297], but not the context in which they encountered the content.[298] At the same time, evidence that are in favor for the opposing side is more often ignored.[289]

The repetition of information can increase confidence in its supposed truthfulness (illusory truth effect).[300] Information that people later find out to be wrong, on the other hand, remain in the memory nonetheless and can have a potentially influencing effect (continued influence effect).[301]

After all, individuals tend to classify familiar content as truthful. The repeated perception of disinformation therefore increases the probability of recipients accepting them as true[302] - in particular, if the information coincides with one's own beliefs.[303] Even the repetition of a statement in the context of countermeasures could strengthen this effect.[304] However, the scientific findings in this area are still unclear.[305]

In addition, there are indications that recipients are not prevented from accessing content by appropriate labels of information as "controversial".[306] Also, content that is not marked as "controversial" or "credible" is considered even more likely to be true.[307] In this respect, the countermeasures explained here seem to have an effect above all if they accompany relevant statements in a comprehensive and reliable way.

Under certain circumstances, labels can even have the opposite effect. The mentioned backfire effect strengthens previous beliefs due to the fact that people are made aware of contradictory evidence.[308] Similarly, information is rejected if it originates from a specific source.[309] "All these biases are compounded by our tendency to believe that we think more impartially than others" (bias blind spot).[310]

A moderately positive effect of corrective information, on the other hand, could be observed in a meta-study which compared the results of 65 individual studies and included the context and nature of the countermeasure in the investigation. It was found that corrections have a moderate effect on beliefs caused by disinformation.[311]

In addition, the study found that corrective information is less popular in the policy context than, for example, in the area of health.[312] Disinformation that deals with real issues (e.g., climate change denial) is also more difficult to correct than those

294 The term describes the effect that the conviction of the users with regard to the correctness of the content was reinforced by the labelling. *Bechmann/O'Loughlin*, 2020, 1 (16) (incl. further ref.); further investigations can be found at *De keersmaecker/Roets*, 2017, 107; *Martel/Mosleh/Rand*, 2017, 17; *Thorson*, 2016, 460; *Walter/Murphy*, 2018, 423; *Brashier/Pennycook/Berinsky/Rand*, 2021, 1; *Pennycook/Rand*, 2017, 1

295 *Sindermann et al*, 2020, 44.

296 *Wolfe/Britt*, 2008, 1

297 *Maier/Richter*, 2013, 151 (152).

298 *Lazer et al*, [fn. 247], 1094 (1095).

299 *Wolfe/Britt*, 2008, 1 (2)

300 *Britt et al*, [fn. 224], 94 (96) (incl. further ref.).

301 *Britt et al*, ibid., 94 (96) (incl. further ref.).

302 *Lazer et al*, [fn. 247], 1094 (1095) (incl. further ref.).

303 S*indermann et al*, 2020, 44 (46).

304 *Lazer et al*, ibid., 1094 (1095).

305 *Lazer et al*, ibid., 1094 (1095) (incl. further ref.).

306 *Bechmann/O'Loughlin*, [fn. 274], 1 (16) (incl. further ref.).

307 *Pennycook/Rand*, 2017, 1 (13).

308 *Bechmann/O'Loughlin*, [fn. 274], 1 (16) (incl. further ref.).

309 *Britt et al*, [fn. 224], 94 (96).

310 *Britt et al*, ibid., 94 (96) (incl. further ref.).

311 „Across 65 individual studies (with a mean sample size of 336.14 and a median of 166) the mean effect size for reduction in post-correction misinformation was moderate, positive, and significant": *Walter/Murphy*, 2018, 423 (432)

312 *Walter/Murphy*, 2018, 423 (428, 436).

dealing with fully fabricated issues (e.g., a fictitious plane crash).[313] In addition, it has been shown that advance warnings are less effective than rebuttals of content. Allegations of lack of coherence (e.g. by providing alternative explanations) also proved to be more effective than fact-checking[314] and hints to lack of credibility.[315]

Accordingly, the time of perception of a correction is also decisive for its effect. A study by Brashier et al. showed that the time at which corrective information is perceived has an influence on how credible headlines from social networks were perceived after a week. Making corrections available after reading the respective heading was more effective than the same information during (by labelling) or before (by prior warnings) exposure.[316]

In summary, it can be stated that the current impact research offers indications for positive effects of corrective information.[317] Contrary effects can be caused by existing views (confirmation bias) or the repeated perception of information. However, these effects represent basic human memory functions - and not necessarily reactions to disinforming content. Therefore, if a legal assessment asks for the pros and cons of such measures to be introduced, the observed positive effects are only offset by the backfire effect. In view of the risk potentials arising from disinformation, measures that make doubts visible thus seem meaningful.[318] Such forms of labelling are also supporting user awareness regarding the circumstance that a statement is debated as true/untrue at all, i.e. whether something is the case or not. With regard to uses of factual assertions to actually communicate an attitude, this alone in fact is a gain for the rationality of the discourse. Nevertheless, it remains the task of science and the platforms to further investigate the effects of labels and tags on the users' side.

Cross-platform Interoperability of Flags, Tags, Warnings and Refutations

If content has been marked with labels or refuting information, such measures are usually limited to the respective platform. This makes it possible to redistribute dubious content across other media outlets and platforms without the doubts already expressed being visible and society's deliberation process being supported. Often, dubious statements are disseminated through embedded links to external sites. Such cases could be approached by the creation of cross-platform labels and interoperability mechanisms. In the case of distribution through embedded links to disinformation, these external links could, for example, be checked against a cross-platform database in order to be able to implement measures across platforms, among other things. Hash value-based procedures, in which a "digital fingerprint" of a statement is created and which can be recognized easily, are also conceivable, but may be easy to circumvent, depending on the design. However, such automated approaches carry the risk of being bypassed by fast-learning and adaptable producers of relevant content.[319] Further starting points could therefore lie in the creation of joint databases of platforms or joint external bodies[320].

4.3.2. Measures in Cases of Statements with a Special Claim to Truth (Journalistic Appearance)

A special case for statements that cannot be clearly falsified are those with a journalistic appearance (see Chapter 2.1.1.). It is irrelevant whether this is a journalistic offer that is subject to professional duties of care or only supposedly journalistic-like looking content. In this respect, pseudo-journalisticofferings also fall within the scope of the following considerations.

The visual appearance or the textual style of such contributions inspires confidence of the recipients that the content conveyed is in line with journalistic and editorial requirements and has therefore been checked for its correctness. From an individual point of view, this information appears to be particularly trustworthy; from a societal point of view, journalistic texts can only fulfil their function of forming individual and public opinion if they provide as correct information as

313 *Walter/Murphy*, ibid., 423 (426) (incl. further ref.).
314 *Walter/Murphy*, 2018, 423 (426). Fact Checking in this context is referred to as: „fact-checking (e.g., determining the veracity of statements regarding political policies)".
315 *Walter/Murphy*, 2018, 423 (434-435).
316 *Brashier et al*, [fn. 240], 1 (2).„We found consistent evidence that the timing of fact-checks matters: „True" and „false" tags that appeared immediately after headlines (debunking) reduced misclassification of headlines 1 week later by 25.3%, compared to an 8.6% reduction when tags appeared during exposure (labeling), and a 6.6% increase (Experiment 1) or 5.7% reduction (Experiment 2) when tags appeared beforehand (prebunking)."
317 Cf. *Pennycook et al*, 2021, 590.
318 *Pennycook/Rand*, 2021, 388.
319 *Caplan et al*, 2018, 1.
320 Regarding advantages of external decision-making bodies in platform governance cf. *Heldt/Dreyer*, 2021, 266. On different forms of soccial media council see *Kettemann/Fertmann*, 2021.

possible. From these two perspectives an increased claim to truth of such information follows. Against this background, pseudo-journalisticcontent proves to be particularly relevant (see Chapter 2.1.2) and particularly threatening with regard to legally protected rights and interests, as they give dubious statements a journalistic appearance and/or use falsified statistics as alleged evidence of their statements. Because of the increased trust of the recipients resulting from the journalistic appearance, potential deceptions can be particularly strong;[321] statements that are underpinned with the help of fake or supposed evidence are proven to be more effective than without evidence. Due to the potentially misleading effect of the content, which could be increased by the exploitation of the increased trust of the recipients, pseudo-journalistic content affects individual and collective (political) decision-making and opinion formation, and ultimately the integrity of elections.

Countermeasures in the area of pseudo-journalistic content have to aim at a counter-factual stabilisation[322] of trust in journalistic content. One approach here is to check the content in question or its producers for compliance with journalistic duties of care when it comes to the production of a statement. This duty of care requires that content be checked as far as possible for its truthfulness before its publication. Since a judgment about truth is often not possible (see Chapter 2.1.2), the obligation to report truthfully must be understood as striving for it. For these reasons, too, checks on compliance with journalistic duties of care are not based on the content itself, but rather on the provision of evidence regarding the journalistic process. Based on these proofs it can be assumed whether news production is compliant with journalistic duties of care, for the sake of which the statement is attributed a special trustworthiness.[323] As addressees of supervisory procedures, offers that fall within the scope of the freedom of the press or broadcasting cannot be considered (see § 109 Sect. 1 S. 4 MStV).

The implementation of respective procedures is demanding very special requirements. For example, guidelines must be made that rule out arbitrary and politically motivated choices of sanctionised criteria. At the level of the supervised subjects, the media's immunity to police action must be taken into account when it comes to enforcement by a supervisory body (see Chapter 3.4); further discussion is still required regarding the appropriate procedures as to how a content provider can or should prove compliance with journalistic duties of care, in particular in areas that are particularly protected under constitutional law, such as journalistic investigation and protection of (confidential) sources.[324]

A legal obligation of platforms to check the compliance of users with journalistic duties of care does not appear possible as an alternative; according to § 10 TMG, host providers are liable for user content only from the time of knowledge. In this respect, obliging private platforms to supervise content production-related requirements constituted a violation of this liability privilege. Should platforms within their private and contractual autonomy stipulate the right to ask content providers with journalistic appearance for proof of compliance with duty of care obligations, they would in principle be free to do so as long as they take into account horizontal effects of basic rights. However, in such cases the platform would then have the power to interprete what would be seen as a journalistic offer and how proof of the necessary duty of care would have to be provided. Both aspects seem problematic in view of the horizontal effects of media freedoms. A solution to this challenge could be seen in state-appointed regulators or self-regulatory bodies developing requirements or guidelines for the scope of "journalistic appearance" and identify best practice guidelines for the substantiation of compliance, while the platforms remain in control over the concrete implementation in their terms of service and internal procedures (see Chapter 4.4.1).

With regard to the measures against a possible violation of duties of care we once again have to refer to the issue of sanctions against content that is based on non-objectively falsifiable content. In this respect, the measures identified in Chapter 4.3 aiming at a lower visibility of dubious content seem to be milder means compared with the traditional instruments of media supervision.[325]

However, the consideration of introducing and assessing a duty of care can not only be applied to journalistical offers, but also to individuals and entities with particularly high relevance or (artificially amplified) wide reach. Such outlets are in a position with communicative power, which could justify an increased expectation towards their responsibility when it

321 *Vargo et al*, 2018, 2028; *Mustafaraj/Metaxas*, 2017.
322 *Luhmann*, 2002, 143-144.
323 Very critical: *Fiedler*, 2021, m.a. 9 et passim.
324 Lent also raises constitutional concerns, *Lent*, 2020, 593.
325 This includes among other things the measures listed in § 109 MStV of prohibition, blockage and retraction as well as revocation.

comes to truthfulness; at least an introduction of increased minimum requirements seem debatable here. The legal concept of linking special responsibilities of communicative power with duties of care can be found in Art. 10 sect. 2 ECHR. The exercise of expression-related fundamental rights, here, "may be subject to such formalities, conditions, restrictions or penalties as are prescribed by law and are necessary in a democratic society". Possible risks for the functioning of democratic society were posed by disinformation in particular if the shown freedom-threatening potentials in the area of decision-making as well as the integrity of the elections would be realized (see Chapter 2.2.2. and 2.2.4.). A restriction of expression-related rights corresponding with Sect. 2 aiming at preserving or supporting the functioning of the basic democratic order therefore does not seem to be far-fetched, at least from the outset. In any case, an intervention would have to be implemented by statutory laws, pursuant to Art. 10 sect. 2 ECHR.

In addition to the above-mentioned requirements for procedures implementing a duty of care for journalistic offerings and applying this to content providers with wide reach, another relevant aspect arises: the determination of a threshold for such increased requirements. The establishment of a minimum value for "communicative power" would first require knowledge about the correlation of range and influence on recipients in order to be able to determine expedient thresholds. In addition, the question arises as to which parameters should be used to measure relevant reach. The number of views of a content or the number of subscribers initially only allows for assuming the reach of an account.[326] Challenges can arise from this situation: the shaping of duties of care must take into account the uncertainty regarding the influence of content with wide reach; the justification of encroachments on fundamental rights – as duties of care would pose on the part of the obligated parties - is all the more presuppositional the greater the uncertainty about the effect that a duty of care aims at minimising. This challenge could be addressed by a "flexible standard of duty of care that varies from case to case", which, depending on the intensity of the intervention in the rights of persons, is influenced by "various care-enhancing or -reducing factors".[327] In view of this necessarily case-by-case assessment, fixed, reach-based thresholds would not be easy to develop. If there is a corresponding political desire to create legal provisions, the legislator has scope for decision-making here.

4.3.3. Measures in Cases of Statements that are Primarily Financially Motivated

As shown above (Chapter 2.1.1), there are forms of disinformation that are less about the ideologically or politically motivated conviction or manipulation of worldviews, but rather about receiving attention and reach primarily for economic reasons. Where, for example, fake or distorted statements attract wide attention, large numbers of recipients and high interaction rates, this content can be monetized via forms of accompanying advertising[328]; high use rates are usually accompanied by increased advertising revenues. Here, entire (pseudo-)journalistic news portals have emerged, which have large numbers of users. From the point of view of risk potentials, untrue statements made exclusively for economic reasons do not differ structurally from other types of disinformation, but the fact that the assertions and information are not created precisely in order to participate in social discourse could lead to a lowered requirement for justification when it comes to legally address such offers.

Against this background proposals have been made that aim at eliminating economic incentives in particular by demonetarizing the respective content.[329] If the goal is to deprive relevant offers and statements of economic opportunities, however, this appears to be a serious interference with the freedom to exercise one's profession and, partially, with the right to an established and ongoing business. The idea of reducing economic incentives for providers of disinforming content presupposes that content has been identified as false – in the context of (court) proceedings or through social negotiation about the truthfulness of a statement. Both procedures are case-by-case, so assessments as to whether only disinformation is being disseminated by an outlet are difficult to prove. The demonetarization of whole portals on the basis of individual incorrect statements appears disproportionate in this respect. For the effective implementation of a general demonetarization of relevant portals, an industry-wide cooperation of advertising networks and services would also be necessary; corresponding attempts at such coordinated behavior, for example in the area of offers that systematically violate copyrights,

326 *Schierbaum*, [fn. 170], 411-412.
327 *Dereje*, VerfBlog v. 05.06.2019, https://verfassungsblog.de/sorgfaltspflichten-auch-fuer-laien-im-netz/; regarding the sliding scale cf. in detail *Schierbaum*, [fn. 170], pp. 409.
328 *Braun/Eklund*, 2019, 1.
329 *Caplan et al*, 2018, 19-20.

have so far failed due to antitrust regulations. A facilitating approach could be that advertisers have the possibility to view the context of the placement of their advertisings in detail as well as to configure advertising context-related restrictions.

However, approaches appear legally permissible and conceivable where platforms for user-generated content stipulated the right not only to label dubious content (see Chapter 4.3.2), but to strip the respective accounts from advertising revenues or to display no advertising at all in this context until further clarification; however, it remains unclear how seriously the advertising industry would pursue this.[330]

4.4. TECHNOLOGY AND DISTRIBUTION-RELATED APPROACHES

As an alternative to content-related measures, it is also possible to discuss countermeasures that focus on the peculiarities of the technical dissemination of disinformation. The spread of (dis)information by new actors is also a socio-technical problem[331], to the extent that disinformation can be disseminated in a targeted and/or far-reaching manner. The uncontrolled, partly viral distribution make up part of the specific risk potentials (see Chapter 2.2.2).

4.4.1. Measures against (Coordinated) Inauthentic Behaviour

"Inauthentic behavior" is a term used by some platforms to describe forms of conduct in which the features and possibilities of the platform are misused to artificially increase the relevance and/or reach of a statement. For example, Facebook defines inauthentic behavior as "use of Facebook or Instagram assets (accounts, pages, groups, or events) with the aim of deceiving other users or Facebook" regarding (a) the identity, purpose or origin of the companies represented by the assets, (b) the popularity of content, (c) the purpose of audiences or communities, or (d) the source or origin of a content. In addition, it is also considered inauthentic behavior when trying to evade the enforcement of community standards through deception.[332]

Some of the forms of inauthentic behavior, in particular the possibility of (artificially) creating reach and relevance as well as the deception about the source of a content, contribute to the potential risks of disinformation (see Chapter 2.2.2 and 2.2.4). Platforms offer technical possibilities that can be used to influence the discourse. This can be achieved in particular through the use of fake account networks and social bots.[333] In this way, disinformation can be communicated on a massive and targeted basis.[334] In addition, a high popularity for topics or people can be feigned[335] and – using the selectio and priorisation logics of social media platforms - artificially created wide reach and relevance of disinformation can be created (see Chapter 2.1.2). This form of inauthentic behavior enables individual actors as well as coordinated networks to take up positions of power within the discourse that are not be communicatively justified (see Chapter 2.2.4 and 3.4), thereby suppressing other points of view.

In any form of technically increased visibility or relevance, however, amplification effects caused by human users also play a role: The spread of disinformation continues to be predominantly caused by humans and not by social bots.[336] Here, technical manipulations can create an important cause for an initially good visibility; in the further course, algorithmic relevance predictions and (further) dissemination by human actors intensify. It is assumed that the cause of human susceptibility to misinformation is that disinformation often provokes strong emotional reactions in recipients.[337] Emotionalizing (false) information usually spreads faster than truthful content, in particular when it concerns political and thus particularly important information for public discourse.[338] By sending, linking, sharing and liking disinformation by users, such content can achieve a wide reach. The circumstance is further reinforced by the platform's own selection and prior-

330 See *Braun/Eklund*, 2019, 1 (17).
331 *Creech*, 2020, 1 (6 et passim).
332 See Item 20 of the Facebook Community Standards, https://www.facebook.com/communitystandards//inauthentic_behavior.
333 *Kind et al*, 2017, pp. 11.
334 *Löber/Roßnagel*, [fn. 71], 154; *Löber/Roßnagel*, 2019, 493.
335 *Thieltges/Hegelich*, 2017, 493 (494-495).
336 *Vosoughi et al*, 2018, 1146 (6).
337 *Shu et al*, [fn. 240], (9 et passim).
338 *Vosoughi et al*, [fn. 23], 1146 (2).

itisation algorithms as well as adoption in classic media reporting. This attention logic of the platform can be exploited[339] by individuals or networks in order to achieve the highest possible spread of disinformation (see Chapter 2.1.1). In addition, hashtags can be used for the targeted amplification of misinformation. For this purpose, often closely networked groups work together to make a hashtag a trend or to "take over" a hashtag.[340] Another form of campaign-like influence is the so-called astroturfing. People who belong to a certain social community use supposedly independent outlets to communicate certain opinions or beliefs and thus give the impression that the widespread opinion originates from a diverse community or a socially widespread view.[341]

In order to reduce the risk potentials arising from these abusive forms of behaviour (see Chapters 2.2.2 to 2.2.4), measures can be considered which do not tie in with the content of a statement but with the patterns of its distribution, especially in view of the factual access possibilities of the platforms to the content and its (further) dissemination. A legal approach to inauthentic behaviour only seems possible in blatant cases of interference with the principle of a right to equal chances in communication. Below such large-scale disinformation campaigns amplified by bot networks and fake accounts, on the other hand, platform governance measures are more likely to come into consideration, especially through contractual regulations in the terms and conditions or terms of use of platforms. In both cases, the specific inauthenticity of a behaviour would have to be defined or at least linked to specific criteria. On the level of statutory regulations, this appears to be more difficult with regard to freedom of expression, since organically generated reach can be seen in particular as a central manifestation of the freedom of expression of many individuals. At the same time, proving non-authentic reach is not trivial. A simple reversal of the burden of proof, whereby platforms would have to prove the authenticity of the generated reach in the event of a complaint, appears to be questionable in terms of EU and constitutional law in view of the aforementioned liability privileges for third-party content and in view of the resulting incentives to sanction a statement in all cases of doubt.

The question of limitations also arises at the level of a (contractual) provision regarding inauthentic behaviour and its definition by the platform: In this case, the power to interpret what is defined as inauthentic behaviour lies exclusively with the platforms themselves. With regard to the horizontal effect of fundamental rights via the general clauses of §§ 242, 307 BGB, the defined procedures must meet certain requirements and in particular be uniformly designed. Terms of service could be seen as inadmissible in particular if they are too vague, or if their application is arbitrary, discriminatory or, due to their design, unduly restricts communication as a whole.[342]

As a mediating approach and with a view to the forms of hybrid governance mentioned earlier, it seems possibl to prescribe abstract mandatory infrastructure measures with which platforms can identify possible abuse scenarios and provide adequate countermeasures. The draft Digital Services Act contains a similar provision with regard to large platforms. According to the draft Art. 26 DSA, very large online platforms must assess systemic risks (also through abuse) and provide for appropriate precautions. Specifically, countermeasures with regard to inauthentic behaviour are presuppositional and range from (automated) recognition to possible sanctions or labelling to the general question of a real-name obligation.

Identification of Social Bots and Other Inauthentic Behaviour

To identify bots, bot networks and organized fake account, it must be assessed whether activities carried out in social networks take place based on human activity or software-based. This can be done by community-based or by automated approaches[343] (see Chapter 4.3.1.1); the automatic recognition of automated accounts and account networks is highly complex[344], though, and can also lead to false positive detections and non-recognitions.[345]

To automatically detect bots and bot networks, behavior-based approaches are particularly important. For example, behavioural patterns can be used for identification (e.g. sharing a large amount of content in a short time or the intervals at

339 "Exploiting vulnerabilities of the media system", *Marwick/Lewis*, 2017, 3.
340 *Agrawal*, 2020, 39.
341 *Zhang et al*, 2013, 1 (1).: „Online astroturfing refers to coordinated campaigns where messages supporting a specific agenda are distributed via the Internet. These messages employ deception to create the appearance of being generated by an independent entity. In other words, astroturfing occurs when people are hired to present certain beliefs or opinions on behalf of their employer through various communication channels. The key component of astroturfing is the creation of false impressions that a particular idea or opinion has widespread support".
342 Kühling, [fn. 96] m.a. 38b (incl. further ref.).
343 *Shu et al*, [fn. 240], (14et passim); Halvani et al, [fn. 236], pp. 111.
344 *Majhi et al*, 2020.
345 *Rauchfleisch/Kaiser*, 2020, 1; *European Parliamentary Research Service (EPRS)*, 2019, 34.

which content is shared).[346] Also, connections within the network can be used for clues (bots often have fewer followers than "real" users).[347] In addition, for example, content-based factors can be used as proxies (indicators can be, for example, certain words, sentences and topics, as well as references to external websites).[348] Similar behavior-based approaches (which, however, are based on other parameters) can also be used to uncover automated bot networks.[349]

However, inauthentic behavior can also be caused by well-connected human networks, which use the selection and prioritization algorithms and the platform's attention logic to spread disinformation. Such organically created reach is within the scope of freedom of expression. And therefore it is difficult to limit by statutory law. Such an organically created scope is in any case difficult to limit by law with regard to freedom of expression, since it reflects actual perceived and explicit relevance of (well-organised) social groups.

In principle, platforms can create own provisions here that link to organically created reach. However, the provisions laid down in the terms of service must meet certain requirements with regard to the horizontal effect of the fundamental rights and must be uniform and nonarbitrary. Due to the possibility of false positives in the identification of organically created range, certain requirements must therefore be met with regard to the procedure. In order to safeguard fundamental rights, such regulations must in particular contain a complaint procedure and should be easily accessible and understandable. Moreover, information should be provided regarding the notion of inauthentic behavior, which behaviors fall under it as well as regarding measures that may follow and ways to complaint about such decisions.

Prohibition and Deletion of Fake Accounts and Social Bots

The most effective way to take action against social bots and the communicative power made possible by them is to prohibit such activities and delete the corresponding accounts.[350] In particular, this could counteract risk potentials for the individual and collective (political) formation of opinion as well as the integrity of elections. With regard to the formation of public opinion, this can specifically reduce potential suppressions of the visibility of certain voices or points of view according to the right of equal chances to communicate.

However, the deletion of social bots constitutes an interference with freedom of expression, which must also be talen into account in the context of prohibitions in the terms of service due to the horizontal effects of fundamental rights. While a bot itself cannot be the holder of fundamental rights, the person behind it can.[351] Article 5 GG (as well as Article 10 ECHR and Article 11 CFR) also covers the free choice of means of communication for one's expression of opinion.[352] Consequently, Art. 5 GG knows no restriction regarding the medium of distribution used for an expression.[353] A general obligation to delete all social bots - whether in statutory law or in terms and conditions of a platform -, regardless of whether they have a harmful effect on the public discourse, would therefore seem disproportionate.[354] A contractual obligation to use real name accounts, on the other hand, is regularly reserved to the platforms; a stutory law obliging users to use real names, on the other hand, is legally difficult (cf. the following paragraph on such obligations).

However, platforms stipulating a deletion of so-called malicious bots seems possible via the terms and conditions; malicious bots are social bots that, inter alia, are used to spread disinformation (see Chapter 2.2.4).[355] To determine when a social bot can be classified as a malicious bot that allows for its deletion, thresholds might be used. Here, for example, a deletion could be initiated depending on how often an automated account has shared disinformation (see Chapter 4.3.2). However, the link to the (regular) sharing or dissemination of disinformation touches on the content-related problems described above. Furthermore, the procedure for such deletions must be transparent, should provide the reasons for the decision and offer complaint mechanisms.

346 *Shu et al*, [fn. 240], 14.
347 *Chu et al*, 2012, 811.
348 *Varol et al*, 2017, 280.
349 *Zipperle*, 2014, 17 (20)
350 *Löber/Roßnagel*, [fn. 71], 186; *Kyza et al*, [fn. 238], 1.
351 *Steinbach*, 2017, 101; *Dankert/Dreyer*, 2017, 73.
352 *Jestaedt*, 2011, m.a. 42.
353 BVerfG, 10. 10.1995 – 1 BvR 1476/91, 1 BvR 1980/91, 1 BvR 102/92, 1 BvR 221/92, BVerfGE 93, 266, 289.
354 In connection with a statutory prohibition cf. *Jansen et al*, 2020, 208; *Löber/Roßnagel*, [fn. 120], 493 (494).
355 *Bader*, 2020, 23.

Labelling of Social Bots

Since a general ban on social bots seems disproportionate except in blatant cases and the deletion of malicious bots would only cover a very limited proportion of disinformation content, labels can be a useful alternative or supplement measure. Where social bots carry a notice regarding their automation, users can easier reflect on the fact that an account is not operated by a human but by a computer program.[356] This, too, could counteract the risk potentials for the individual and collective (political) formation of opinion as well as the integrity of elections, since respective labels enable more critical assessments of the statements made.

However, it is again to be asked whether the interference with freedom of expression is justified by a general labelling obligation regarding bots. This consideration must include the potentially harmful effects of abusive uses of social bots regarding public discourse and free public and, in particular, political decision-making (see Chapter 2.2). Against the background that such obligations merely reveal that communication is automated, disclosure duties for automated accounts appear to be proportionate and thus compatible with freedom of expression.[357] Thus, a statutory rule can also be considered here. Such regulations already exist in Germany; according to §§ 18 Sect. 3, 93 sect. 4 MStV, owners of automated accounts and providers of media intermediaries are obliged to identify automatically created statements as such, provided that the user accounts have been "made available for use by natural persons according to their appearance".

Also with regard to the labelling of social bots (by the platform), the procedure must be transparent and understandable for users. In addition, a complaint procedure, as well as easily accessible and understandable information, must exist for cases where accounts have been incorrectly marked as bots.

Real Name Policies

Another systemic approach to counteract the problem of inauthentic behavior is the introduction of a clear name policy on social media platforms, which would oblige users to use and prove their full first and last name for their profile in order to avoid misuse. This can strengthen the responsibility of users and develop a deterrent effect to spread disinformation. However, obligations to provide real names is questionable from a constitutional point of view, as such a provisions may deprive persons of the only opportunity to express certain information or views without risking state, political or social repercussions. Before this background, it is not possible for the state to impose such an obligation to use real names in view of freedom of expression. However, platforms could be entitled to opt for real name policies if they need actual names, for example, for the operation of the business.[358] Admittedly, § 13 sect. 6 TMG contains an obligation for telemedia providers to enable users to use their services anonymously or pseudonymously, but its applicability to online intermediaries based in EU or EEA states has not yet been conclusively clarified: § 1 Sect. 5 BDSG (Federal Data Protection Act) stipulates that the general data protection law is not applicable if the registered office of a responsible body is located in an EU or EEA state. There, (only) the respective national data protection laws apply.[359] In the case of Facebook, whose EU establishment is based in Ireland, an administrative court (VG Schleswig) has already ruled that German data protection law and thus § 13 Sect. 6 TMG does not apply in such cases.[360]

Labelling of Individuals and Institutions with Special Claims to Truth

The verification or labelling of accounts can serve also as a basis for a positive visualization or prioritisation of particularly credible sources and subsequently lead to an increased perception of information that can be seen as particularly trustworthy. Such approaches would apply in particular to outlets working in a journalistic capacity and adhering to professional ethics. By verifying an account and proving compliance with journalistic standards, corresponding accounts and statements could be labelled in a specific way. As currently implemented verifications only confirm identity, this does not lead to more trust in the truthfulness of the disseminated content per se. Moreover, the criterion of a person's or organisation's fame, which is a prerequisite for verification, does not provide any information about the trustworthiness of the information disseminated via the respective accounts. A starting point could therefore be to extend verifications of journalistic outlets by the factor of trustworthiness. On a voluntary basis, they could provide proof of their particular trustworthiness, for example by documenting that they have joined a Press Council. Consequently, if the accounts were

356 *Löber/Roßnagel*, 2019, 493 (494).
357 *Dankert/Dreyer*, 2017, 73 (78).
358 OLG München, GRUR-Prax 2021, 30; regarding the issues cf. *Kluge*, 2017, 230; *Caspar*, 2015, 233.
359 See VG Schleswig, ZD 2013, 245.
360 See VG Schleswig, ZD 2013, 245.

labelled accordingly by the platform, it would be obvious to the users that they can trust the contents of the accounts with regard to the truthfulness of the information disseminated.

However, legally obliging platforms to offer such verification procedures seems problematic. In particular, the criteria of verification should not be determined by the legislator for reasons of a possible state-based discrimination of mass media statements. The possibility of indirect state influence on the determination of preferential content must also be ruled out. Therefore, the platforms themselves come into consideration. If they not only create the criteria for verification, but also have the authority to interpret whether the criteria are fulfilled, the platforms could gain a powerful position in the assessment of what is particularly credible. In addition, they could give certain views a special position of power within the public discourse. Cooperation with external, independent verification bodies could counteract such risks, though. In addition, hybrid governance approaches also seem conceivable here, where the legislator provides the (opinion-agnostic) guiding criteria for types of positive prioritisation but leaves their concretisation to bodies that are independent both from the state and the platform providers, and who can be measured against the legal requirements.[361]

4.4.2. Special Cases: Transparency and Duties of Identification for Political Ads

A special possibility of spreading disinformation is political microtargeting. Here, political advertising messages are played out to users according to their interests and preferences in a targeted manner.[362] The intention behind this is the assumption that users can be efficiently reached and convinced of the disseminated political positions in this way.[363] True and untrue claims, but also - as is often the case in political discourse - opinions, world views or attitudes can be communicated via such channels. If the distributed content is disinformation, there is a risk of reinforcing or manipulating people in their political positions based on false assumptions.[364]

Regardless of the truth of a political ad, targeted content has the structurally similar problem that in these cases political claims are made that have social relevance, but are removed from public discourse, including the negotiation of the truth of the statements made. Political microtargeting therefore also has the potential to interfere with several freedoms discussed in Chapter 2.2.2 and 2.2.4, in particular with regard to individual freedom of opinion and decision-making. If the influence on individuals through microtargeting would actually have a manipulative effect, not only an interference with the individual right to form a political opinion would be assumed, but also with regard to the formation of a public will would be affected due to the interaction between these two aspects (see Chapter2.2.4.). A special quality of intervention arises when the technical distribution possibilities of the platform are exploited in such a way that the advertisements are only visible to selected user groups and there is no possibility to debate the content in the public sphere or to let it become the subject of social discourse. Such a "discourse circumvention" show structural comparability with content that is disseminated immediately before an election act, where there no longer remains sufficient time for a discursive engagement with the content (see Chapter 4.2).[365] Microtargeting, too, can impair the free decision-making process protected by Article 38 GG in the run-up to an election. Whether an inadmissible influence on elections is exceeded on top by the exercise of coercion or pressure has to be assessed on a case-by-case basis cases and has to consider the content of the ad.

Hence, microtargeting represents a form of dissemination of disinformation, which could reinforce the effects of disinformation through the possibilities of targeted addressing and the circumvention of discourse. Establishing legal countermeasures in this area, however, is challenging: this is due in particular to the yet uncertain effects of microtargeting on the part of the recipients. As long as negative and positive effects cannot be empirically proven, it is difficult to balance the affected legal interests[366], especially in an area as highly relevant for opinion-forming as political communication. The terms "political communication" or "political advertising" already have enormous potential for hardly foreseeable overspill effects in core areas of legitimate socio-political disputes. Since parties are *supposed* to convince citizens of their political programmes and plans - this is one of their central functions - large-scale restrictions regarding the form of addressing potential voters appear problematic. However, in order to make parties accountable for statements which are removed

361 As far as the prioritization of offers on user interfaces is concerned, see § 84 Sect. 5 MStV.
362 *Zuiderveen Borgesius et al*, 2018, 82 (82).
363 *Haller/Kruschinski*, 2020, 519.
364 See *Zuiderveen Borgesius et al*, [fn. 338], 82; *Bayer*, 2020, 1.
365 Legislators have identified this risk in the analogue world, see § 32 BWahlG.
366 The rights of freedom are contrasted with the right of political parties to participate in the formation of the will of the people, vgl. Art. 21 p. 1 S. 1 GG.

from public discourse, measures for the general perceptibility of political advertising and all advertisements shown[367] appear to be particularly useful. An approach already implemented by Facebook, for example, could also be to reserve the right to take measures against advertising with falsifiable content.[368] Regarding the requirements for the procedure for falsifiability of content, see Chapter 4.3.1.2. In addition, obligations and self-commitments, especially regarding election campaign communication, can be seen as a way forward.[369] In addition, there are possibilities to limit the criteria for targeting selections of addressees through the negative exclusion or positive restriction of targeting-relevant segments.

As long as the described research deficit and the boundless scope of application of provisions for "political communication" persist, measures that promote cooperation between the platform and researchers should be taken into account in particular (see Chapter 4.1).

4.4.3. Limitation of the Distribution of Individual Statements

Restrictions of Sharing Functions

A potential distribution-related restriction could be a general limitation of sharing or redistributing content. Fixed limits of (re-)distribution can be seen as a starting point here. Accordingly, it is proposed to limit sharing per person (e.g. one person can forward a piece of content to a maximum of 3/5/10 people) or per contribution (e.g. a specific content can be shared a maximum of 100/1000 times). It would also be conceivable to implement limitations that focus on the number of (re-)transmissions per unit of time. Such approaches could in particular counteract risk potentials for the individual and collective (political) formation of will and opinion as well as the integrity of elections. However, instruments limiting the dissemination of content beyond a certain reach – if it were implemented by statutory law – is not convincing on several levels. The determination of a limit value regarding reach is challenging: It would have to be based on actual effects of the contents, which are hardly comprehensible and depend on the individual individual case (see Chapter 4.3.2 regarding this problem). Even more serious is the fact that such measures constitute an interference with the freedom of expression of Art. 5 GG of all users who want to redistribute content, plus it intervenes in the central communication logic of social media platforms, whose providers are also carriers of fundamental rights. Even those users who are denied the display of the contents are infringed in their rights guaranteed under Art. 5 GG. The limited content would be taken away from communication across publics and could not become part of the societal discourse processes.

The implementation of a sharing-related limitation by a platform itself, for example by setting limit values in its terms of use, opens up a wide range of questions with regard to the actual binding of platforms to fundamental rights. In this case, it would be conceivable to review the community standards by means of a control of terms and conditions, which in turn can have fundamental rights balancing questions as its object due to the horizontal effects of fundamental rights. In the context of a possible assessment, the above-mentioned legal positions would then have to be brought into a balance. The procedure-related requirements would then encounter the same challenges for determining a limit value for reach as already mentioned above.

Acquisition of Artificial Reach

If actors on social media platforms generate their reach in an artificial way - i.e. by buying followers, likes or shares - the platform could be obliged to impose respective disclosure provisions on them. Transparent notices would make it clear how many followers were purchased and what funds were used to finance the purchase. Expecially the latter can provide information about the context and motives of a content and its distribution. Labelling these accounts and/or their content can be seen as a basis for enabling discourse. In practice, current terms of use of major platform providers already contain clauses in which they reserve the right to delete profiles with artificially purchased reach. A general statutory prohibition of the purchase of followers or (alleged) reach, on the other hand, seems problematic in view of freedom of expression, which might also encompass purchases in view of the freedom of choice of the means of communication. However, as explained, a red line can be crossed where, through corresponding purchases, reach and visibility are achieved that categorically impair other views in their perceptibility - or factually displace them altogether.

367 These include identity verification and authorisation for the placement of political advertising, disclosure of its funding, and the introduction of "ad archives", cf. *Leerssen et al*, 2019.

368 Cf. No 13 of the Facebook Advertising Policies, https://www.facebook.com/policies/ads/.

369 Civil society actors are beginning to build up pressure here, cf. for example the "Leitfaden für Digitale Demokratie" (Guide to Digital Democracy) published in the run-up to the Bundestag election campaign, which is aimed primarily at political parties. Disclosure: The Leibniz Institute for Media Research | Hans Bredow Institute has been involved in ist drafting. https://campaign-watch.de/

4.5. OFFICIAL ANNOUNCEMENTS

It also seems possible to make disinforming content a subject of discussion through official announcements. The Singaporean government, for example, reserves the right to use official announcements, which must be displayed on social media platforms, to point out circulating false reports (such as in May 2021, when the government ordered the display of an announcement with which it countered information about a CoVid19 variant that allegedly originated in Singapore[370]).[371] Official announcements can have the advantage that they not only provide a dubious statement with a reference to its untruthfulness or doubtfulness, but can also potentially counteract variations of the original disinformation. As already explained (see Chapter 3.3.2), such forms of information-based state action would have to be limited to exceptional cases in which there is a concrete and immediate danger. In addition, legal obligations to publish state information entail the rsik of exploitation of such an infrastructure for state propaganda; the possibility of counteracting disinformation through official announcements also creates the option of discrediting certain persons and opinions publicly or of propagating a certain government view. Here, the state can theoretically secure a position of supremacy in the discourse that is not justified communicatively. Where such announcements are made outside of current catastrophic or dangerous situations, the dimension of negative freedom of information has to be taken into account, as it offers protection against state indoctrination, especially through propaganda.[372]

4.6. EDUCATIONAL MEASURES

If citizens recognize and understand types and forms of disinformation channels and means of their dissemination as well as possible effects, they can critically question and reflect on the content in question. Users also evaluate disinformation more often when they think (more) analytically.[373] It is therefore obvious that there is a widespread demand for educational initiatives and awareness-raising campaigns among media users.[374] If media education is to be understood as a countermeasure to disinformation, it is also important to acknowledge that it does not represent a measure against disinforming content specifically, but has an indirect and long-term effect. In particular, it can also contribute to improving[375] the underlying causes of disinformation.[376] As a result, education is not to be denied its relevance as a suitable measure; in the context of the other comparable specific measures discussed in this report, it is only briefly mentioned here. In any case, there are no legal objections to the promotion of media education programs; in particular, no interference with the protected legal rights (see Chapter 3) is apparent.

4.7. SYNOPSIS: OPTIONS FOR ACTION

The study reveals a number of possible approaches to countering the risk dimensions of disinformation. Four major areas of action that can be identified are: (a) measures to improve regulatory knowledge, (b) measures in cases of objectively falsifiable content, (c) measures in cases of doubts about non-falsifiable statements, and (d) measures targeting technological- and distribution-related aspects independently of the content. The following overview (Table 4) summarizes possible countermeasures, the requirements regarding their design and the actors called upon to implement them.

370 Yahoo News, 21.05.2021, COVID: POFMA orders issued over false 'Singapore variant' claims, https://sg.news.yahoo.com/covid-pofma-orders-issued-singapore-variant-comments-030441241.html.
371 *Jie*, 2020.
372 *Grabenwarter*, 2020, m.a. 1018-1019.
373 *Sindermann et al.*, 2020, 44 (46).
374 *Kyza et al*, [fn. 238], 1; *Roozenbeek/Linden*, 2019, 570; *Britt et al*, [fn. 224], 94. With a particular focus on educating influencers: *Bechmann*, 2020, 1. On the limits of the effect of education in relation to media consumers' consent to misinformation, cf. *Hameleers*, 2020, 1.
375 Regarding correlations of larger societal contexts and the disinformation-related resilience of individual states cf. *Humprecht et al*, 2020, 493.
376 *Sängerlaub et al*, 2018, pp. 94.

Table 4: Options for action, their requirements and relevant implementing actors

Measure	Requirements and risks	Design	Actor
Area of measures A: Measures to improve regulatory knowledge			
Information / disclosure obligations; access rights	- Accuracy of the data is not fully verifiable - Access rights can enable validation	- Congruent and comparable report structure and data structures - Provision of country- specific data	National or EU legislator
Area of measures B: Measures in cases of objectively falsifiable statements			
Legal prohibitions	- Legal case-by-case decision regarding the trueness of a claim required - Great and imminent danger for legally protected rights necessary (life and health; public safety; free elections)	- Court-like, uniform proceedings for each individual case	Nationaler Gesetzgeber
Reservation of rights to delete or block content	- Accessory to legal prohibitions, or voluntary reservations in terms of use (in case of the latter a threat to legally protected rights is not necessary)	- Uniform proceedings; duty to give reasons - Option for persons affected to object - Transparent procedures	Platforms; national or EU legislator
Downranking/ De-prioritization	- Legal provisions only in case of a threat to legally protected rights - Reservation of rights to apply such measures on voluntary basis in terms of use (in this case a threat to legally protected rights is not necessary) - Potential threat to freedom of expression due to excessive deprioritization - Recognizability of deprioritization for affected persons is limited	- Uniform proceedings; duty to give reasons - Notification of affected persons; option for these persons to object - Transparent procedures	National or EU legislator in case of threat to legal rights Platforms; national or EU legislator with legal (minimum)- requirements for the procedure
Tagging of falsified content	- Content remains available for discourse; no "censorship" - Risk of misuse in community-based approaches	- Uniform proceedings; duty to give reasons - Option for persons affected to object - Transparent procedures	National or EU legislator Platforms; national or EU legislator with legal minimum)- requirements for the procedure

Area of measures C: Measures in cases of doubts about not or not completely falsifiable statements			
Reporting / flagging mechanisms	- Possible requirement for subsequent fact checking procedures - User-based flagging has high requirements and is prone to misuse	- Uniform processes - Trusted flaggers as an option for more reliable reports and starting point for prioritized processing	Platforms
Fact checking procedures	- Dominant position of fact checking bodies - High costs with low scalability - Requirement for follow-up measures (esp. tagging, counter-speech based measures)	- Easily accessible and understandable information regarding the criteria, procedure and follow-up measures - Implementation by independent institutions / bodies - Uniform processes; cross-platform bodies as an option - Option for persons affected to object with external bodies - Transparent and auditable processes - Exceptions for small and non-profit platforms	Platforms; national or EU legislator with legal (minimum)-requirements for the procedure
Labelling / tagging	- Risk of misuse in context of community-based approaches - Design options: notice or warning, with or without reference to refuting information - also possible at the level of whole accounts or profiles (strong intervention) - Knowledge of effects of labelling / tagging is still limited	- Uniform processes - Option for persons affected to object - Transparent and auditable processes - Accompanying research on effects necessary	Platforms; national or EU legislator with legal (minimum)-requirements for the procedure
Addition of contradicting information / debunking	- Risk of misuse in context of community-based approaches	- Uniform processes - Option for persons affected to object - Transparent and auditable processes - Accompanying research on effects necessary	Platforms; national or EU legislator with legal (minimum)-requirements for the procedure

Special case: Measures regarding statements with special claims of truth			
Obligation to exercise journalistic duties of care in cases of journalistic appearance	- High conformity with expectations of journalistic functions - Danger of misuse a duty of care obligations by the state - Potential interpretative power of platforms regarding what is considered journalistic appearance - Legal provisions that oblige platforms to monitor / control are not possible	- Exclusion of state intervention, e.g. through arbitrary or politically motivated selection of targets - Procedures must be independent from the state - Guidelines for the distinction of journalistic appearance necessary - Development of criteria for providing evidence required	National legislator, implementation by regulators Independent implementation by platforms
Obligation to exercise journalistic duties of care in cases of accounts with high relevance or wide reach	- Difficulty in determining the threshold for duties of care to be applicable - Necessity of a case-by-case assessment	- Development of flexible standards of duties of care varying from case to case	National legislator, implementation by regulators Independent implementation by platforms
Special case: Measures regarding financially motivated statements			
Capping of economic incentives / demonetarisation	- Infringement with the freedom to choose and carry out one's career as well as the right to an established and operating business - Demonetisation of entire outlets or profiles based on few false statements is disproportionate	- Potential anti-trust issues regarding cross-provider agreements - Options for advertisers to select advertising contexts and to subsequent review	Advertising industry (advertising networks, advertisers)
Area of measures D: Measures targeting technological- and distribution-related aspects			
Prohibition of social bots	- General legal prohibition disproportionate; restriction to blatant cases of interference with the right to equal chances to communicate - Challenging legal criteria since evidence of automation is difficult to supply; general issue of the burden of proof - Permission and limitation by platforms possible	- Uniform processes - Option for persons affected to object - Transparent and auditable processes	Platforms; national or EU legislator with legal (minimum) requirements for platform concretisations and for procedures;

Labelling of social bots	- Obligations to disclose automated communication is proportionate	When labelling is carried out by platforms: - Uniform processes - Option for persons affected to object - Transparent and auditable processes	Platforms; national or EU legislator with legal (minimum)- requirements for the procedure
Real name policies	- Legal obligation to provide real name seems questionable with regard to fundamental rights	- Obligation to provide real name on basis of platform terms possible	Platforms
Positive labelling of persons / institutions with valid special claims to truth	- Legal obligation only possible in cases where state independent processes have identified actors who can be granted a positive label - Otherwise: Strong position of the state or the platform regarding the decision-making power over criteria and verification processes - Indirect potentials for misuse (e.g., through positive labelling of only certain outlets)	- Cooperation with external, independent verification bodies - (Non-binding) state guidelines regarding possible criteria feasible - External review of the concretisation by platforms	Platforms; national or EU legislator with guiding criteria and, if necessary, framework requirements for the design of positive labels
Special case: Measures regarding political microtargeting			
Obligations regarding transparency and mandatory identification for political advertisers	- Legal requirements possible, but interference with key functions of political parties - Challenge: definition of "political advertising"	- Focus on visibility and public discourse regarding booked political ads - Alternative approach: Limitation of selectable segments in political advertising - Self-commitments by parties as a less restrictive measure	National or EU legislator; platforms; political parties
Special case: Measures regarding technical limitations of the distribution of statements			
Restrictions of sharing functionalities	- Strong interference in freedom of expression; legal provisions questionable with regard to human rights	- Implementation and limitation by platform possible - Counterproductive with regard to social discourse	Platforms
Prohibition or labelling of cases of buying artificial reach	- Legal provisions seem problematic; - Tagging by platforms is legitimate	- Uniform processes - Option for persons affected to object	Platforms; national or EU legislator with legal (minimum)- requirements for the procedure

Area of measures E: Official statements			
Official Statements	- Potential for misuse: Possibility of discrediting specific persons / groups / opinions by propagating a certain governmental viewpoint	- Restriction to exceptional cases of high and imminent danger	EU or national legislator

Source: Own illustration.

The overview in Table 4 shows that legislative options for regulation of disinformation are limited to a few and severe types of disinformation. Where classic legal instruments such as legal prohibtions reach their limits in cases of possible disinformation, further possibilities exist for platforms to shape the rules of communication within their offerings, especially based on their contractual autonomy. However, these margins to shape their platform governance do not apply indefinitely. The (new) power of platforms in shaping public and private communication is bound to take human rights into account when implementing internal processes. However, this is not an insight that applies to disinformation specifically, but a part of the general legal debate in the field of regulating platforms with user generated content ("platform governance" or "governance of platform governance").

In light of the countermeasures that platforms can introduce to counter disinformation the importance of minimum legal requirements regarding such procedures has become clear. In this field, there is an opportunity for EU and national legislators to develop principles, guidelines, and benchmarks that apply in cases where platforms decide to provide certain countermeasures. By doing so, legislators are able to safeguard the respect for fundamental rights once platforms implement content-related processes. Article 12 of the draft Digital Services Act, for example, seems to be a first step in such direction with regard to the formulation of terms of service. In the medium term the DSA might become a regulatory platform for formulations of legal guidelines which are then given concrete forms and are implemented by private parties, open for their subsequent monitoring and review by socially accountable institutions and bodies. The evaluation exercise of the Code of Practice on Disinformation shows hints that the EU is increasingly thinking in terms of forms of co-regulation in this area; this would be consistent with the line of thinking in this study. The form of societal self-efficacy by using new actors could be guaranteed by different organizational or procedural provisions. One possible form for such bodies can be seen in media regulators that are independent from the state and consist of pluralistic decision-making bodies, such as e.g. the German state media authorities.

5. SUMMARY

(1) **Scientific definitions** of disinformation focus on the untruthfulness of a statement and the speaker's intention to mislead. From a legal perspective both criteria are difficult to determine and therefore are **rarely suitable for regulatory debates**. For the study, we therefore use the following working definition: "Disinformation describes utterances,
(1) the truth of which can be doubted with good reason,
(2) which can easily be disseminated and shared,
(3) which due to the person making the statement or due to their design claim to be truthful from an objective recipient's perspective, and
(4) which impair legally protected rights and goods."

(2) Such statements can be further analysed by applying a number of **dimensions**, enabling a more **differentiated assessment** and policy discourse. These dimensions are
- the type of statement,
- the context of the statement,
- the structure of the actor making the statement,
- the motive for making the statement and underlying incentives,
- the degree of its potential public visibility,
- a recognisable intention to mislead or deceive in a given case, and
- pivotally, its potential risk for legally protected rights and goods (with the dimensions of the immediacy of danger, the probability of a violation, the intensity of the violation, and the importance of the legally protected rights concerned).

(3) In order to answer the question to what extent there is a need for action against the spread of disinformation and what options are available, we first need to **identify legally protected rights that might be affected** by disinformation. We identify the following rights (in some cases additional rights may be affected):

1. On the individual and group level
   - autonomy,
   - the freedom of political decision-making and opinion-forming,
   - right to free election,
   - freedom of expression,
   - freedom of information,
   - the general right of personality,
   - the right to unimpaired personal development, and
   - the rights to life and health.

2. At the societal level
   - the freedom of public opinion formation,
   - equal opportunities to communicate,
   - diversity of opinion,
   - the democratic formation of will and the integrity of elections,
   - trust in democratic institutions,
   - the construction of societal reality and social cohesion, as well as
   - public safety and order and public health.

(4) Contrasting the affected interests and legally protected rights with the current legal framework – using Germany as an example – shows that **individual and group-related legal rights are essentially protected** against dangers emanating from (online) communication. The effectiveness of this protection may be subject to criticism, for example with regard to the possibilities of the legal system to deal with the quantity and rate of dissemination of potentially harmful communications on platforms. However, this problem exists in relation to all content that endangers legally protected rights and is not limited to content that involves dangers due to their untruthfulness.

(5) The latter represent a problem category when it comes to the possibilities of platform-specific legal reactions in order to make protection more effective. But if the violation of a legal right specifically rests upon the untruthfulness of, for example, an allegation about a person, an adequate solution to **the legal dispute presupposes that - depending on the burden of proof - it must be possible to prove the truth or untruthfulness**. Procedures such as the NetzDG require platforms to take this decision, although they do not have court-like proceedings to find the truth. An adequate conflict resolution is not only about determining the facts of the case, but also about the question of whether there is a factual claim at all and, if so, what the exact content of the statement encompasses. The incentives for platforms to delete content due to the obligation to check statements lead to the question whether such regulatory concepts can be designed in ways that would be in accordance with fundamental human rights.

(6) With regard to societal interests and rights, specific regulations in Germany only exist for few legally protected interests such as public peace and the counterfactual stabilisation of trust in journalistic content through provisions regarding journalistic duties of care (§ 19 MStV). Any regulatory options in this area **face the problem that in order to assess true/untrue statements, certain bodies would have to be able to determine the truth**. Apart from adversarial proceedings in independent courts, however, **declaring the truth is not a task of the state, but a social process of communicative construction in which substantiation and doubt play an important role**. At least temporarily, shared understandings of "what is the case" emerge in such processes.

(7) The **functioning of such discursive processes can only be guaranteed to a very limited extent by state measures**. If trust in actors or individuals with a particular role in the construction of reality erodes, this has an impact on the democratic self-understanding of a society. This is the case, for example, when political-strategic claims are currently being made in the USA that the presidential election was manipulated, although courts have rejected this. Political actors should be aware of the fundamental danger of changing the practices of reality construction. The preservation of political culture appears to be the central factor in this case. Self-commitments, especially regarding campaign communication, might be an option. However, areas remain where the state cannot enable this social process, but might at least be able to support it through legal frameworks.

(8) Across all countermeasures examined, it appears that **traditional forms of regulation are either not permissible for such discourse-supporting approaches, or are not helpful, or do not appear to be feasible in practice**. One reason for this is that private actors would regularly be responsible for their implementation, who themselves are entitled to basic freedoms when it comes to shaping their offers and contractual conditions. Here, **in addition to classic forms of self-regulation, new forms of "hybrid governance" are needed**, where state regulation and platforms' own areas of governance are intertwined. State-based control of communication can rarely contribute to solving the problems associated with disinformation; or only at the price of endangering the very freedom that it aims to protect.

(9) Against this background, **only a few paths appear to be expedient when it comes to containing problems of disinformation by regulation**, and only a few of them use the criterion of (un)truth. Apart from the reasons mentioned, the difference between true/untrue is related to the type of statement. However, this does not always correspond to the actual use of language: currently, language seems to be used in a way that underlines a political statement by offensive denial of facts (e.g. "I don't care that it is disproved, the election was stolen"). Here, the underlying problem is not solved by checking the truth of the statement. Based on the findings of the study the following paths could be pursued further:

- Legal measures based on untruthfulness can be considered (only) if there is a **high probability of immediate danger to the highest individual legal rights such as life and physical integrity**. This includes statements that may mobilise a lynch mob or factual allegations that might directly cause self-harm. In these cases, after balancing the legally protected interests, statements might exceptionally first be deleted and then checked for their untruthfulness in legal proceedings. In these cases, public discourse cannot fulfil its function because it might be too late by then.

- Another area of legal provisions can target untrue statements made **in direct temporal proximity to a democratic election**. In these cases, too, society is deprived of the possibility to negotiate the truthfulness of the statement. Where voters can be manipulated in such a way, society's interest in the freedom to vote outweighs the rights of the person making that statement.

- **Measures that make doubts about a statement visible and check the statement** (tagging or labelling, fact-checking procedures, notices and warnings as well as combinations of these) cannot be introduced on a mandatory basis by law with regard to legal utterances, but are **subject to the voluntary measures of the platforms**. However, expectations and ideas can be exchanged through cooperation between the state and platforms as well as among the platforms themselves. If platform providers introduce such measures and procedures, the (EU or national) legislator should provide a legal framework that safeguards fundamental rights regarding these procedures. This might include the obligation to transparently provide information in the terms of use regarding the existence of such procedures and their decision-making processes, possible sanctions and the rights of those affected by these decisions.

- The current legal framework for the counterfactual stabilisation of **trust in journalistic content requires compliance with journalistic duties of care**. This is **one way of linking statements claiming truthfulness with their increased obligations to seek truth**. If a sliding scale of duties of care is applied, **non-journalistic actors with high relevance for opinion formation** (such as influencers or activists) might also be covered by such obligations. Such measures have the additional advantage that they contribute to stabilising expectations of certain types of offerings. In the light of the variety of content provided on platforms, this is a significant aspect for both users and providers. Due to the potential for abuse of such a regulatory approach the aspects of selection, proof of compliance and sanctions must be designed in a particularly careful, transparent and comprehensible manner to prevent any suspicion of forms of supervision that target specific opinions. With regard to necessary sanctions in the case of violations of duties of care, forms of labelling can be seen as milder means compared to injunctions or deletion orders.

- Access to or duties to disclose platform data, usually a rather generic demand in policy discussions, makes specific sense in the area of disinformation: only then can society learn which discourse-oriented measures to make doubt visible (e.g. by labelling or counter-speech) show impact. Regulation could work towards this, ideally in such a way that an established procedure or data broker does not have to negotiate the conditions of data access for each individual case.

(10) **Governing inauthentic behaviour on platforms** can also be a way to slow down the viral spread of disinformation. However, this approach points to a general problem of platform regulation that can only be touched upon here: **a platform has a legitimate interest in defining what it considers to be authentic communication by its users**. In this area of "hybrid" public/private governance, new forms of interaction between state and private norm-setting bodies seem effective. In this case, this could consist of government-appointed regulators formulating requirements for authenticity from the perspective of societal interests. However, it would then be up to the platforms to implement this in a more detailed way in their terms of service, and then control their implementation. This instrument is only apparently content-neutral, as certain actors can be recognised through specific patterns of sharing or liking, for example.

(11) A key future challenge in dealing with disinformation is the search for **possibilities of cross-platform measures that aim at making doubt and fact checking results visible**. If the validity of a statement is disputed based on valid evidence and arguments on one platform, interoperable forms of making this doubt visible can help prevent the same statement from remaining unquestioned on other platforms.

(12) Measures that are not specifically related to disinformation but nevertheless compensate for it can be seen in systematically improving the information literacy of children, adolescents and adults, in positively labelling providers who are committed to journalistic duties of care as well as in discussing forms of (more) disinformation-sensitive reporting by traditional media and journalistic outlets.

(13) Governance measures designed to reduce the potential risks of disinformation should not distract from the fact that the increased occurrence of disinformation may have deeper societal causes. If disinformation is a symptom, states and societies can only solve the issue in the long run if they also address such underlying causes in parallel.

REFERENCES

Agrawal, Neelesh, The Digital Challenges to Democracy: Social Media and New Information Paradigms, 2020.

Allcott, Hunt / Gentzkow, Matthew, Social Media and Fake News in the 2016 Election, Journal of Economic Perspectives 2017, 211–236.

Armbrüster, Christian, § 134, in: Säcker, Franz-Jürgen / Rixecker, Roland / Oetker, Hartmut / Limperg, Bettina (Hrsg.), Münchener Kommentar zum Bürgerlichen Gesetzbuch, 8. Auflage, München 2018.

Ash, Timothy Garton / Gorwa, Robert / Metaxa, Danaë, GLASNOST! Nine ways Facebook can make itself a better forum for free speech and democracy, An Oxford Standard Report 2018, 1–28.

Bader, Katarina, Einleitung, in: Steinebach, Martin / Bader, Katarina / Rinsdorf, Lars / Krämer, Nicole / Roßnagel, Alexander (Hrsg.), Desinformation aufdecken und bekämpfen. Interdisziplinäre Ansätze gegen Desinformationskampagnen und für Meinungspluralität, Einleitung, Baden-Baden 2020.

Bader, Katarina / Jansen, Carolin / Rinsdorf, Lars, Jenseits der Fakten: Deutschsprachige Fake News aus Sicht der Journalistik, in: Steinebach, Martin / Bader, Katarina / Rinsdorf, Lars / Krämer, Nicole / Roßnagel, Alexander (Hrsg.), Desinformation aufdecken und bekämpfen Interdisziplinäre Ansätze gegen Desinformationskampagnen und für Meinungspluralität, Jenseits der Fakten: Deutschsprachige Fake News aus Sicht der Journalistik, Baden-Baden 2020.

Bail, Christopher A. / Guay, Brian / Maloney, Emily / Combs, Aidan / Hillygus, D. Sunshine / Merhout, Friedolin / Freelon, Deen / Volfovsky, Alexander, Assessing the Russian Internet Research Agency's impact on the political attitudes and behaviors of American Twitter users in late 2017, Proc Natl Acad Sci USA 2020, 243–250.

Bakir, Vian / McStay, Andrew, Fake News and The Economy of Emotions: Problems, causes, solutions, Digital Journalism 2018, 154–175.

Barbu, Oana, Advertising, Microtargeting and Social Media, Procedia - Social and Behavioral Sciences 2014, 44–49.

Barua, Zapan / Barua, Sajib / Aktar, Salma / Kabir, Najma / Li, Mingze, Effects of misinformation on COVID-19 individual responses and recommendations for resilience of disastrous consequences of misinformation, Progress in Disaster Science 2020, 1–9.

Bateman, Jon / Thompson, Natalie / Smith, Victoria, How Social Media Platforms' Community Standards Address Influence Operations, 2021.

Bayer, Judit, Double harm to voters: data-driven micro-targeting and democratic public discourse, Internet Policy Review 2020, 1–17.

Bayer, Judit / Bitiukova, Natalija / Bárd, Petra / Szakács, Judit / Alemanno, Alberto / Uszkiewicz, Erik / Carrera, Sergio / Vosyliute, Lina / Guérin, Julia, Disinformation and propaganda – impact on the functioning of the rule of law in the EU and its Member States, Brüssel 2019.

Bechmann, Anja, Tackling Disinformation and Infodemics Demands Media Policy Changes, Digital Journalism 2020, 1–9.

Bechmann, Anja / O'Loughlin, Ben, Democracy & Disinformation: A Turn in the Debate, KVAB Thinkers' Report 2020, 1–37.

Bentele, Günter, Der Wahrheits- und Wahrhaftigkeitsanspruch in einer Ethik der öffentlichen Kommunikation, in: Schicha, Christian / Stapf, Ingrid / Sell, Saskia (Hrsg.), Medien und Wahrheit: Medienethische Perspektiven auf Desinformation, Lügen und „Fake News", Der Wahrheits- und Wahrhaftigkeitsanspruch in einer Ethik der öffentlichen Kommunikation, Baden-Baden 2021.

Berger, Peter L. / Luckmann, Thomas, Die gesellschaftliche Konstruktion der Wirklichkeit: eine Theorie der Wissenssoziologie, 26. Auflage, Frankfurt am Main 2016.

Blake-Turner, Christopher, Fake news, relevant alternatives, and the degradation of our epistemic environment, Inquiry 2020, 1–21.

Bliss, Nadya / Bradley, Elizabeth / Garland, Joshua / Menczer, Filippo / Ruston, Scott W. / Starbird, Kate / Wiggins, Chris, An Agenda for Disinformation Research, 2020, 5.

Böckenförde, Ernst-Wolfgang, Recht, Staat, Freiheit: Studien zur Rechtsphilosophie, Staatstheorie und Verfassungsgeschichte, Frankfurt am Main 1991.

Böckenförde, Ernst Wolfgang, Staat, Verfassung, Demokratie: Studien zur Verfassungstheorie und zum Verfassungsrecht, 2. Auflage, Frankfurt am Main 1992.

Böckenförde, Ernst-Wolfgang, § 24 Demokratie als Verfassungsprinzip, in: Isensee, Josef / Kirchhof, Paul (Hrsg.), Handbuch des Staatsrechts der Bundesrepublik Deutschland, § 24 Demokratie als Verfassungsprinzip, 3., völlig neubearbeitete und erw. Aufl, Heidelberg 2003.

Brashier, Nadia M. / Pennycook, Gordon / Berinsky, Adam J. / Rand, David G., Timing matters when correcting fake news, Proc Natl Acad Sci USA 2021, 1–3.

Braun, Joshua A. / Eklund, Jessica L., Fake News, Real Money: Ad Tech Platforms, Profit-Driven Hoaxes, and the Business of Journalism, Digital Journalism 2019, 1–21.

Britt, M. Anne / Rouet, Jean-François / Blaum, Dylan / Millis, Keith, A Reasoned Approach to Dealing with Fake News, Policy Insights from the Behavioral and Brain Sciences 2019, 94–101.

Brown, Étienne, Propaganda, Misinformation, and the Epistemic Value of Democracy, Critical Review 2018, 194–218.

Buchheim, Johannes, Rechtlicher Richtigkeitsschutz, Der Staat 2020, 159–194.

Camebridge Consultants, Use of AI in Online Content Moderation, 2019, 1–82.

Caplan, Robyn / Hanson, Lauren / Donovan, Joan, Dead Reckoning. Navigating Content Moderation After "Fake News", 2018, 1–38.

Caspar, Johannes, Klarnamenpflicht versus Recht auf pseudonyme Nutzung, ZRP 2015, 233–236.

Castillo, Carlos, Big Crisis Data: Social Media in Disasters and Time-Critical Situations, 2016.

Chambers, Simone, Truth, Deliberative Democracy, and the Virtues of Accuracy: Is Fake News Destroying the Public Sphere?, Political Studies 2021, 147–163.

Chu, Zi / Gianvecchio, Steven / Wang, Haining / Jajodia, Sushil, Detecting Automation of Twitter Accounts: Are You a Human, Bot, or Cyborg?, IEEE Trans. Dependable and Secure Comput. 2012, 811–824.

Claussen, Victor, Fighting hate speech and fake news. The Network Enforcement Act (NetzDG) in Germany in the context of European legislation, Medialaws 2018, 110–136.

Colliver, Chloe, Cracking the Code: An Evaluation of the EU Code of Practice on Disinformation, 2020.

Cornils, Matthias, Art. 10 EMRK & Art.11 GRCh in: Gersdorf, Hubertus / Paal, Boris P. / Paal, Boris P. (Hrsg.), BeckOK Informations- und Medienrecht, 31. Auflage, München.

Cottee, Simon, Can Facebook Really Drive Violence?, The Atlantic 2018.

Craufurd Smith, Rachael, Fake news, French Law and democratic legitimacy: lessons for the United Kingdom?, Journal of Media Law 2019, 52–81.

Creech, Brian, Fake news and the discursive construction of technology companies' social power, Media, Culture & Society 2020, 1–17.

Dankert, Kevin / Dreyer, Stephan, Social Bots – Grenzenloser Einfluss auf den Meinungsbildungsprozess? Eine verfassungsrechtliche und einfachgesetzliche Einordnung, K&R 2017, 73–78.

De keersmaecker, Jonas / Roets, Arne, 'Fake news': Incorrect, but hard to correct. The role of cognitive ability on the impact of false information on social impressions, Intelligence 2017, 107–110.

Dereje, Laura, Sorgfaltspflichten auch für Laien im Netz!, https://verfassungsblog.de/sorgfaltspflichten-auch-fuer-laien-im-netz/.

Di Fabio, Udo, Art.2 Abs.1, in: Maunz, Theodor / Dürig, Günter / Herzog, Roman (Hrsg.), Grundgesetz: Kommentar, München 2020.

Douek, Evelyn, Self-regulation of content moderation as an answer to the special problems of speech regulation, Aegis Paper Series 2019, 1–28.

Dreyer, Stephan / Heldt, Amelié, Algorithmische Selektion und Privatheit. Aufmerksamkeitssteuerung durch Social-Media-Plattformen als Autonomieeingriff?, in: Berger, Franz X. / Derametz, Anne / Henning, Martin / Mitchell, Alix (Hrsg.): Verantwortung in digitalen Kulturen. Privatheit im Geflecht von Medien, Recht und Gesellschaft, Baden-Baden 2021.

Duffy, Andrew / Ling, Rich, The Gift of News: Phatic News Sharing on Social Media for Social Cohesion, Journalism Studies 2020, 72–87.

Egelhofer, Jana Laura / Lecheler, Sophie, Fake news as a two-dimensional phenomenon: a framework and research agenda, Annals of the International Communication Association 2019, 97–116.

Elias, Norbert, Symboltheorie, Frankfurt am Main 2001.

Elsaß, Lennart / Labusga, Jan-Hendrik / Tichy, Rolf, Löschungen und Sperrungen von Beiträgen und Nutzerprofilen durch die Betreiber sozialer Netzwerke, CR 2017, 234–241.

Engeler, Malte, Meinungsfreiheit: Warum Facebook (zu Recht) nicht an Grundrechte gebunden ist, in: Krone, Jan (Hrsg.), Medienwandel kompakt 2017-2019, Meinungsfreiheit: Warum Facebook (zu Recht) nicht an Grundrechte gebunden ist, Wiesbaden 2019.

Eser, Albin, § 108a, in: Schönke, Adolf / Schröder, Horst (Hrsg.), Strafgesetzbuch: Kommentar, 30., neu bearbeitete Auflage, München 2019.

European Parliamentary Research Service (EPRS) (Hrsg.), Automated tackling of disinformation, Brüssel 2019.

Fallis, Don, The varieties of disinformation, in: Floridi, Luciano / Illari, Phyllis (Hrsg.), The philosophy of information quality, The varieties of disinformation, Cham 2014.

Fallis, Don, What Is Disinformation?, Library Trends 2015, 401–426.

Fertmann, Martin / Potthast, Keno C., Digitale timeouts für Trump: Der Anfang vom Ende der privilegierten Behandlung von Amtsinhaber*innen durch soziale Netzwerke?, https://www.juwiss.de/05-2021/.

Fiedler, Christoph, § 109 MStV, in: Gersdorf, Hubertus / Paal, Boris P. (Hrsg.), BeckOK Informations- und Medienrecht, 31. Edition, München 2021.

Gaozhao, Dongfang, Flagging Fake News on Social Media: An Experimental Study of Media Consumers' Identification of Fake News, Government Information Quarterly 2021, 1–24.

Gelfert, Axel, Fake News: A Definition, Informal Logic 2018, 84–117.

Gorwa, Robert / Ash, Timothy Garton, Democratic Transparency in the Platform Society, in: Persily, Nathaniel / Tucker, Joshua A. (Hrsg.), Social Media and Democracy: The State of the Field, Prospects for Reform, Democratic Transparency in the Platform Society, 1, 2020.

Gostomzyk, Tobias, Grundrechtsträgerschaft für soziale Netzwerke? Der Anwendungsbereich des Art. 19 Abs. 3 GG, in: Eifert, Martin / Gostomzyk, Tobias (Hrsg.), Netzwerkrecht, Grundrechtsträgerschaft für soziale Netzwerke? Der Anwendungsbereich des Art. 19 Abs. 3 GG, Baden-Baden 2018.

Grabenwarter, Christoph, Art. 5 Abs. 1 Abs. 2, in: Maunz, Theodor / Dürig, Günter / Herzog, Roman (Hrsg.), Grundgesetz: Kommentar, München 2020.

Grinberg, Nir / Joseph, Kenneth / Friedland, Lisa / Swire-Thompson, Briony / Lazer, David, Fake news on Twitter during the 2016 U.S. presidential election, Science 2019, 374–378.

Guess, Andrew / Nagler, Jonathan / Tucker, Joshua, Less than you think: Prevalence and predictors of fake news dissemination on Facebook, Sci. Adv. 2019, 1–8.

Habermas, Jürgen, Wahrheitstheorien, in: Fahrenbach, Helmut / Schulz, Walter (Hrsg.), Wirklichkeit und Reflexion: Walter Schulz zum 60. Geburtstag, Wahrheitstheorien, Pfullingen 1973.

Habermas, Jürgen, Strukturwandel der Öffentlichkeit: Untersuchungen zu einer Kategorie der bürgerlichen Gesellschaft, Frankfurt am Main 1990.

Haller, André / Kruschinski, Simon, Politisches Microtargeting. Eine normative Analyse von datenbasierten Strategien gezielter Wähler_innenansprache, ComSoc 2020, 519–530.

Halvani, Oren / Heereman von Zuydtwyck, Wendy / Herfert, Michael / Kreutzer, Michael Kreutzer / Liu, Huajian / Simo Fhom, Hervais-Clemence / Steinebach, Martin / Vogel, Inna / Wolf, Ruben / Yannikos, York / Zmudzinski, Sascha, Automatisierte Erkennung von Desinformationen, in: Steinebach, Martin / Bader, Katarina / Rinsdorf, Lars / Krämer, Nicole / Roßnagel, Alexander (Hrsg.), Desinformation aufdecken und bekämpfen. Interdisziplinäre Ansätze gegen Desinformationskampagnen und für Meinungspluralität, Automatisierte Erkennung von Desinformationen, Baden-Baden 2020.

Hameleers, Michael, Separating truth from lies: comparing the effects of news media literacy interventions and fact-checkers in response to political misinformation in the US and Netherlands, Information, Communication & Society 2020, 1–17.

Hameleers, Michael / Powell, Thomas E. / Van Der Meer, Toni G.L.A. / Bos, Lieke, A Picture Paints a Thousand Lies? The Effects and Mechanisms of Multimodal Disinformation and Rebuttals Disseminated via Social Media, Political Communication 2020, 281–301.

Hartl, Korbinian, Suchmaschinen, Algorithmen und Meinungsmacht, Wiesbaden 2017.

Hasebrink, Uwe / Schulz, Wolfgang / Sprenger, Regina / Rzadkowski, Nora, Macht als Wirkungspotenzial. Zur Bedeutung der Medienwirkungsforschung für die Bestimmung vorherrschender Meinungsmacht, Berlin 2009.

Heidtke, Aron, Meinungsbildung und Medienintermediäre, Baden-Baden 2020.

Heins, Markus / Lefeldt, Stefanie, Medienstaatsvertrag: Journalistische Sorgfaltspflichten für Influencer*innen, MMR 2021, 126–130.

Heldt, Amélie / Dreyer, Stephan, Competent Third Parties and Content Moderation on Platforms: Potentials of Independent Decision-Making Bodies From A Governance Structure Perspective, Journal of Information Policy 2021, 266–300.

Heldt, Amélie / Dreyer, Stephan / Schulz, Wolfgang / Seipp, Theresa Josephine, Normative Leitbilder der Europäischen Medienordnung: Leitvorstellungen und rechtliche Anforderungen an die Governance für eine demokratische Öffentlichkeit, 2021, 1–31.

Helm, Rebecca K / Nasu, Hitoshi, Regulatory Responses to 'Fake News' and Freedom of Expression: Normative and Empirical Evaluation, Human Rights Law Review 2021, 302–328.

Hey, C. / Jacob, K./ Volkery, A., REACH als Beispiel für hybride Formen von Steuerung und Governance, in: Schuppert, G.F./Zürn M. (Hrsg.), Governance in einer sich wandelnden Welt. VS Verlag für Sozialwissenschaften, 2008.

Hindelang, Steffen, Freiheit und Kommunikation: Zur verfassungsrechtlichen Sicherung kommunikativer Selbstbestimmung in einer vernetzten Gesellschaft, Berlin, Heidelberg 2019.

Hoffmann-Riem, Wolfgang, Art. 5, in: Denninger, Erhard / Hoffmann-Riem, Wolfgang / Schneider, Hans Peter (Hrsg.), Kommentar zum Grundgesetz für die Bundesrepublik Deutschland (AK-GG), Art. 5, 3, Neuwied 2001.

Hoffmann-Riem, Wolfgang, Medienregulierung als objektiv-rechtlicher Grundrechtsauftrag, M&K 2002, 175–194.

Hoffmann-Riem, Wolfgang, Regulierungswissen in der Regulierung, in: Bora, Alfons / Henkel, Anna / Reinhardt, Carsten (Hrsg.), Wissensregulierung und Regulierungswissen, Regulierungswissen in der Regulierung, Weilerswist 2014.

Högden, Birte / Krämer, Nicole / Meinert, Judith / Schaewitz, Leonie, Wirkung und Bekämpfung von Desinformation aus medienpsychologischer Sicht, in: Steinebach, Martin / Bader, Katarina / Rinsdorf, Lars / Krämer, Nicole / Roßnagel, Alexander (Hrsg.), Desinformation aufdecken und bekämpfen. Interdisziplinäre Ansätze gegen Desinformationskampagnen und für Meinungspluralität, Wirkung und Bekämpfung von Desinformation aus medienpsychologischer Sicht, Baden-Baden 2020.

Holzer, Stefanie / Sengl, Michael, Quelle gut, alles gut? Glaubwürdigkeitsbeurteilung im digitalen Raum, in: Hohlfeld, Ralf / Harnischmacher, Michael / Heinke, Elfi / Lehner, Lea / Sengl, Michael (Hrsg.), Fake News und Desinformation: Herausforderungen für die vernetzte Gesellschaft und die empirische Forschung, Quelle gut, alles gut? Glaubwürdigkeitsbeurteilung im digitalen Raum, Baden-Baden 2020.

Humprecht, Edda, Where 'fake news' flourishes: a comparison across four Western democracies, Information, Communication & Society 2019, 1973–1988.

Humprecht, Edda / Esser, Frank / Van Aelst, Peter, Resilience to Online Disinformation: A Framework for Cross-National Comparative Research, The International Journal of Press/Politics 2020, 493–516.

Jack, Caroline, Lexicon of Lies: Terms for Problematic Information, 2017, 1–20.

Jankowicz, Nina / Pierson, Shannon, Freedom and Fakes: A Comparative Exploration of Countering Disinformation and Protecting Free Expression, 2020, 1–33.

Jansen, Carolin / Johannes, Paul Christopher / Krämer, Nicole / Kreutzer, Michael Kreutzer / Löber, Lena Isabell / Rinsdorf, Lars / Roßnagel, Alexander / Schaewitz, Leonie / Wolf, Ruben / Yannikos, York / Zmudzinski, Sascha, Handlungsempfehlungen, in: Steinebach, Martin / Bader, Katarina / Rinsdorf, Lars / Krämer, Nicole / Roßnagel, Alexander (Hrsg.), Desinformation aufdecken und bekämpfen. Interdisziplinäre Ansätze gegen Desinformationskampagnen und für Meinungspluralität, Handlungsempfehlungen, Baden-Baden 2020.

Jestaedt, Matthias, § 102 Meinungsfreiheit, in: Merten, Detlef / Papier, Hans-Jürgen (Hrsg.), Handbuch der Grundrechte in Deutschland und Europa, § 102 Meinungsfreiheit, 2011.

Jie, Woo Jun, Between Surveillance and Security: The Protection from Online Falsehoods and Manipulation Bill (POFMA), 2020.

Jung, Heike, Über die Wahrheit und ihre institutionellen Garanten, JZ 2009, 1129.

Jungherr, Andreas / Schroeder, Ralph, Disinformation and the Structural Transformations of the Public Arena: Addressing the Actual Challenges to Democracy, Social Media + Society 2021, 1–13.

Kalla, Joshua L. / Broockman, David E., The Minimal Persuasive Effects of Campaign Contact in General Elections: Evidence from 49 Field Experiments, Am Polit Sci Rev 2018, 148–166.

Kapantai, Eleni / Christopoulou, Androniki / Berberidis, Christos / Peristeras, Vassilios, A systematic literature review on disinformation: Toward a unified taxonomical framework, New Media & Society 2021, 1301–1326.

Katsirea, Irini, "Fake news": reconsidering the value of untruthful expression in the face of regulatory uncertainty, Journal of Media Law 2018, 159–188.

Keller, Daphne / Leersen, Paddy, Facts and Where to Find Them: Empirical Research on Internet Platforms and Content Moderation, in: Persily, Nathaniel / Tucker, Joshua A. (Hrsg.), Social Media and Democracy: The State of the Field, Prospects for Reform, Facts and Where to Find Them: Empirical Research on Internet Platforms and Content Moderation, 1, 2020.

Kettemann, Matthias C. / Fertmann, Martin, Die Demokratie plattformfest machen, 2021.

Kettemann, Matthias C. / Schulz, Wolfgang, Setting Rules for 2.7 Billion: a (First) Look into Facebook's Norm-Making System; Results of a Pilot Study, Working Papers of the Hans-Bredow-Institut | Works in Progress 2020.

Kind, Sonja / Jetzke, Tobias / Weide, Sebastian / Ehrenberg-Silies, Simone / Bovenschulte, Marc, Social Bots - Die potenziellen Meinungsmacher, 2017, 1–80.

Kischel, Uwe, in: Epping, Volker / Hillgruber, Christian (Hrsg.), BeckOK Grundgesetz, 3. Auflage, München 2020.

KJM, Schwerpunktanalyse 2020: „Alternative Medien und Influencer als Multiplikatoren von Hass, Desinformation und Verschwörungstheorien", 2020.

Klein, Friedrich, Art. 41, in: Maunz, Theodor / Dürig, Günter / Herzog, Roman (Hrsg.), Grundgesetz: Kommentar, München 2020.

Kluge, Steffen, Klarnamenspflicht bei Facebook – Rechtliche Grenzen und Möglichkeiten, K&R 2017, 230–236.

Köhler, Helmut, Zur Neuvermessung der Tatbestände der unzumutbaren Belästigung, WRP 2017, 253–262.

Kušen, Ema / Strembeck, Mark, Politics, sentiments, and misinformation: An analysis of the Twitter discussion on the 2016 Austrian Presidential Elections, Online Social Networks and Media 2018, 37–50.

Kühling, Jürgen, Art. 5 GG, in: Gersdorf, Hubertus / Paal, Boris P. / Paal, Boris P. (Hrsg.), BeckOK Informations- und Medienrecht, 31. Auflage, München 2018.

Kyza, Eleni A / Varda, Christiana / Panos, Dionysis / Karageorgiou, Melina / Komendantova, Nadejda / Perfumi, Serena Coppolino / Shah, Syed Iftikhar Husain / Hosseini, Akram Sadat, Combating misinformation online: re-imagining social media for policy-making, IPR 2020, 1–24.

Landesanstalt für Medien NRW (Hrsg.), Was ist Desinformation? Betrachtungen aus sechs wissenschaftlichen Perspektiven, Düsseldorf 2020.

Lazer, David M. J. / Baum, Matthew A. / Benkler, Yochai / Berinsky, Adam J. / Greenhill, Kelly M. / Menczer, Filippo / Metzger, Miriam J. / Nyhan, Brendan / Pennycook, Gordon / Rothschild, David / Schudson, Michael / Sloman, Steven A. / Sunstein, Cass R. / Thorson, Emily A. / Watts, Duncan J. / Zittrain, Jonathan L., The science of fake news, Science 2018, 1094–1096.

Leerssen, Paddy / Ausloos, Jef / Zarouali, Brahim / Helberger, Natali / Vreese, Claes H. de, Platform ad archives: promises and pitfalls, Internet Policy Review 2019.

Lent, Wolfgang, Paradigmenwechsel bei den publizistischen Sorgfaltspflichten im Online-Journalismus – Zur Neuregelung des § 19 Medienstaatsvertrag, ZUM 2020, 593–600.

Lewandowsky, Stephan / Ecker, Ullrich K.H. / Cook, John, Beyond Misinformation: Understanding and Coping with the "Post-Truth" Era, Journal of Applied Research in Memory and Cognition 2017, 353–369.

Löber, Lena Isabell / Roßnagel, Alexander, Kennzeichnung von Social Bots. Transparenzpflichten zum Schutz integrer Kommunikation, MMR 2019, 493–498.

Löber, Lena Isabell / Roßnagel, Alexander, Desinformation aus der Perspektive des Rechts, in: Steinebach, Martin / Bader, Katarina / Rinsdorf, Lars / Krämer, Nicole / Roßnagel, Alexander (Hrsg.), Desinformation aufdecken und bekämpfen Interdisziplinäre Ansätze gegen Desinformationskampagnen und für Meinungspluralität, Desinformation aus der Perspektive des Rechts, Baden-Baden 2020.

Luhmann, Niklas, Erkenntnis als Konstruktion, Bern 1988.

Luhmann, Niklas, Die Realität der Massenmedien, 1996.

Luhmann, Niklas, Das Recht der Gesellschaft, 1. Auflage, [Nachdr.], Frankfurt am Main 2002.

Luhmann, Niklas, Soziale Systeme: Grundriß einer allgemeinen Theorie, 17. Auflage, Frankfurt am Main 2018.

Madiega, Tambiama, Reform of the EU liability regime for online intermediaries: background on the forthcoming digital services act: in depth analysis., Luxemburg 2020.

Mafi-Gudarzi, Nima, Desinformation: Herausforderung für die wehrhafte Demokratie, ZRP 2019, 65–68.

Maier, Johanna / Richter, Tobias, Text Belief Consistency Effects in the Comprehension of Multiple Texts With Conflicting Information, Cognition and Instruction 2013, 151–175.

Majhi, Surjya Prasad / Swain, Santosh Kumar / Pattnaik, Prasant Kumar, Issues of Bot Network Detection and Protection, in: Mallick, Pradeep Kumar / Balas, Valentina Emilia / Bhoi, Akash Kumar / Chae, Gyoo-Soo (Hrsg.), Cognitive Informatics and Soft Computing, Issues of Bot Network Detection and Protection, Singapore 2020.

Marret, Christophe, The impact of social media on elections, NUPRI Working Paper Nr. 4 2020, 1–19.

Marsch, Nikolaus, Das europäische Datenschutzgrundrecht: Grundlagen, Dimensionen, Verflechtungen, Tübingen 2018.

Martel, Cameron / Mosleh, Mohsen / Rand, David G, 6 You're Definitely Wrong, Maybe: Correction Style Has Minimal Effect on 7 Corrections of Misinformation Online, 17.

Marwick, Alice / Lewis, Rebecca, Media Manipulation and Disinformation Online, 2017.

Masing, Johannes, Meinungsfreiheit und Schutz der verfassungsrechtlichen Ordnung, JZ 2012, 585.

Mast, Tobias, Staatsinformationsqualität: De- und Rekonstruktion des verfassungsgerichtlichen Leitbilds öffentlicher staatlicher Informationstätigkeit und der entsprechenden Gebote, Berlin 2020.

Mayen, Thomas, Über die mittelbare Grundrechtsbindung Privater in Zeiten des Einflusses sozialer Netzwerke auf die öffentliche Kommunikation, ZHR 2018, 1–7.

McKay, Spencer / Tenove, Chris, Disinformation as a Threat to Deliberative Democracy, Political Research Quarterly 2020, 1–15.

Möller, Judith / Hameleers, Michael / Ferreau, Frederik, Typen von Desinformation und Misinformation. Verschiedene Formen von Desinformation und ihre Verbreitung aus kommunikationswissenschaftlicher und rechtswissenschaftlicher Perspektive. Gutachten im Auftrag der Gremienvorsitzendenkonferenz der Landesmedienanstalten (GVK)., Berlin 2020.

Monsees, Linda, 'A war against truth' - understanding the fake news controversy, Critical Studies on Security 2020, 1–14.

Morgan, Susan, Fake news, disinformation, manipulation and online tactics to undermine democracy, Journal of Cyber Policy 2018, 39–43.

Müller, Karsten / Schwarz, Carlo, Fanning the Flames of Hate: Social Media and Hate Crime, SSRN Journal 2017, 1–46.

Müller, Karsten / Schwarz, Carlo, Making America Hate Again? Twitter and Hate Crime Under Trump, SSRN Journal 2018, 1–51.

Mustafaraj, Eni / Metaxas, Panagiotis Takis, The Fake News Spreading Plague: Was it Preventable?, arXiv:1703.06988 [cs] 2017, http://arxiv.org/abs/1703.06988.

Nickerson, Raymond S., Confirmation Bias: A Ubiquitous Phenomenon in Many Guises, Review of General Psychology 1998, 175–220.

Oreskes, Naomi, Why trust science?, Princeton, NJ 2019.

Pearson, George, Sources on social media: Information context collapse and volume of content as predictors of source blindness, New Media & Society 2021, 1181–1199.

Pennycook, Gordon / Epstein, Ziv / Mosleh, Mohsen / Arechar, Antonio A. / Eckles, Dean / Rand, David G., Shifting attention to accuracy can reduce misinformation online, Nature 2021, 590–595.

Pennycook, Gordon / Rand, David G., Assessing the Effect of "Disputed" Warnings and Source Salience on Perceptions of Fake News Accuracy, SSRN Journal 2017, 1–33.

Pennycook, Gordon / Rand, David G., The Psychology of Fake News, Trends in Cognitive Sciences 2021, 388–402.

Pörksen, Bernhard, Die Deregulierung des Wahrheitsmarktes. Von der Macht der Desinformation im digitalen Zeitalter, in: Blamberger, Günter / Freimuth, Axel / Strohschneider, Peter / Weduwen, Karena (Hrsg.), Vom Umgang mit Fakten: Antworten aus Natur-, Sozial- und Geisteswissenschaften, Die Deregulierung des Wahrheitsmarktes. Von der Macht der Desinformation im digitalen Zeitalter, Paderborn 2018.

Pöttker, Horst, Wahrheit und Wahrhaftigkeit, Communicatio Socialis 2017, 85–89.

Puschmann, Cornelius, Technische Faktoren bei der Verbreitung propagandistischer Inhalte im Internet und den sozialen Medien, in: Schmitt, Josephine B. / Ernst, Julian / Rieger, Diana / Roth, Hans-Joachim (Hrsg.), Propaganda und Prävention, Technische Faktoren bei der Verbreitung propagandistischer Inhalte im Internet und den sozialen Medien, Wiesbaden 2020.

Rader, Emilee / Gray, Rebecca, Understanding User Beliefs About Algorithmic Curation in the Facebook News Feed, Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems 2015, 173–182.

Rauchfleisch, Adrian / Kaiser, Jonas, The False positive problem of automatic bot detection in social science research, PLoS ONE 2020, 1–20.

Reinhardt, Jörn, Letzter Vorhang für @realdonaldtrump., VerfBlog 2021.

Reuter, Christian / Hartwig, Katrin / Kirchner, Jan / Schlegel, Noah, Fake News Perception in Germany: A Representative Study of People's Attitudes and Approaches to Counteract Disinformation, 2019.

Rini, Regina, Fake News and Partisan Epistemology, Kennedy Institute of Ethics Journal 2017, E-43-E-64.

Roozenbeek, Jon / van der Linden, Sander, The fake news game: actively inoculating against the risk of misinformation, Journal of Risk Research 2019, 570–580.

Ross, Björn / Pilz, Laura / Cabrera, Benjamin / Brachten, Florian / Neubaum, German / Stieglitz, Stefan, Are social bots a real threat? An agent-based model of the spiral of silence to analyse the impact of manipulative actors in social networks, European Journal of Information Systems 2019, 394–412.

Saliger, Frank, Kann und soll das Recht die Lüge verbieten?, in: Depenheuer, Otto (Hrsg.), Recht und Lüge, Kann und soll das Recht die Lüge verbieten?, Münster 2002.

Samuel-Azran, Tal / Hayat, Tsahi, Online news recommendations credibility: The tie is mightier than the source, Comunicar 2019, 71–80.

Sängerlaub, Alexander, Feuerwehr ohne Wasser? Möglichkeiten und Grenzen des Fact-Checkings als Mittel gegen Desinformation, 2018.

Sängerlaub, Alexander / Meier, Miriam / Rühl, Wolf-Dieter, Fakten statt Fakes: Das Phänomen „Fake News". Verursacher, Verbreitungswege und Wirkungen von Fake News im Bundestagswahlkampf 2017, Berlin 2018.

Schaal, Gary S., Hybride Diskurs-Beeinflussung. Angriffe auf die demokratische Öffentlichkeit durch ausländische Propaganda, in: Russ-Mohl (Hrsg.), Streitlust und Streitkunst: Diskurs als Essenz der Demokratie, Hybride Diskurs-Beeinflussung. Angriffe auf die demokratische Öffentlichkeit durch ausländische Propaganda, Köln 2020.

Schäfer, Jürgen, § 130, in: Joecks, Wolfgang / Miebach, Klaus / Germany (Hrsg.), Münchener Kommentar zum Strafgesetzbuch, 3. Auflage, München 2017.

Schemmer, Franz, in: Epping, Volker / Hillgruber, Christian (Hrsg.), BeckOK Grundgesetz, 46. Edition Auflage, München.

Scherer, Helmut, Massenmedien, Meinungsklima und Einstellung: eine Untersuchung zur Theorie der Schweigespirale, Opladen 1990.

Scherer, Inge, Das Chamäleon der Belästigung - Unterschiedliche Bedeutungen eines Zentralbegriffs des UWG, WRP 2017, 891–896.

Schierbaum, Laura, Sorgfaltspflichten von professionellen Journalisten und Laienjournalisten im Internet: zugleich ein Beitrag zur rechtlichen Einordnung einer neuen Publikationskultur, Baden-Baden 2016.

Schink, Alan, Verschwörungstheorie und Konspiration: Ethnographische Untersuchungen zur Konspirationskultur, Wiesbaden 2020.

Schulz, Wolfgang, Gewährleistung kommunikativer Chancengleichheit als Freiheitsverwirklichung, 1. Auflage, Baden-Baden 1998.

Schulz, Wolfgang / Dreyer, Stephan, Governance von Informations-Intermediären - Herausforderungen und Lösungsansätze, 2020.

Seifert, Josef, De Veritate - Über die Wahrheit: 1: Wahrheit und Person. 2: Der Streit um die Wahrheit, Berlin/Boston 2013.

Shenkman, Carey / Thakur, Dhanaraj / Llansó, Emma, Do You See What I See? Capabilities and Limits of Automated Multimedia Content Analysis, 2021, 1–62.

Shu, Kai / Bhattacharjee, Amrita / Alatawi, Faisal / Nazer, Tahora / Ding, Kaize / Karami, Mansooreh / Liu, Huan, Combating Disinformation in a Social Media Age, arXiv:2007.07388 [cs] 2020, http://arxiv.org/abs/2007.07388.

Sindermann, Cornelia / Cooper, Andrew / Montag, Christian, A short review on susceptibility to falling for fake political news, Current Opinion in Psychology 2020, 44–48.

Skirbekk, Gunnar (Hrsg.), Wahrheitstheorien: eine Auswahl aus den Diskussionen über Wahrheit im 20. Jahrhundert, 12. Auflage, Frankfurt am Main 2016.

Spindler, Gerald / Schmitz, Peter / Liesching, Marc (Hrsg.), Telemediengesetz: mit Netzwerkdurchsetzungsgesetz : Kommentar, 2. Auflage, München 2018.

Stapf, Ingrid, Wahrheit als moralische Grundlage der Zivilgesellschaft, in: Filipovic, Andreas (Hrsg.), Medien- und Zivilgesellschaft, Wahrheit als moralische Grundlage der Zivilgesellschaft, Weinheim 2012.

Stapf, Ingrid, „Fake News" als eine (mögliche) Frage der Wahrheit? Medienethische Perspektiven auf Wahrheit im Kontext der Digitalisierung, in: Schicha, Christian / Stapf, Ingrid / Sell, Saskia (Hrsg.), Medien und Wahrheit: Medienethische Perspektiven auf Desinformation, Lügen und „Fake News", „Fake News" als eine (mögliche) Frage der Wahrheit? Medienethische Perspektiven auf Wahrheit im Kontext der Digitalisierung, Baden-Baden 2021.

Stark, Birgit / Stegmann, Daniel / Magin, Melanie / Jürgens, Pascal, Are Algorithms a Threat to Democracy? The Rise of Intermediaries: A Challenge for Public Discourse, 2020, 1–69.

Steinbach, Armin, Social Bots im Wahlkampf, ZRP 2017, 101–105.

Steinebach, Martin / Bader, Katarina / Rinsdorf, Lars / Krämer, Nicole / Roßnagel, Alexander (Hrsg.), Desinformation aufdecken und bekämpfen. Interdisziplinäre Ansätze gegen Desinformationskampagnen und für Meinungspluralität, Baden-Baden 2020.

Susser, Daniel / Roessler, Beate / Nissenbaum, Helen, Technology, autonomy, and manipulation, Internet Policy Review 2019.

Tandoc, Edson C. / Lim, Zheng Wei / Ling, Richard, Defining "Fake News": A typology of scholarly definitions, Digital Journalism 2018, 137–153.

Thieltges, Andree / Hegelich, Simon, Manipulation in sozialen Netzwerken, ZfP 2017, 493–512.

Thorson, Emily, Belief Echoes: The Persistent Effects of Corrected Misinformation, Political Communication 2016, 460–480.

Turcilo, Lejla / Obrenovic, Mladen, Misinformation, Disinformation, Malinformation: Causes, Trends, and Their Influence on Democracy, 2020, 1–38.

Ungern-Sternberg, Antje, Demokratische Meinungsbildung und künstliche Intelligenz, in: Unger, Sebastian / Ungern-Sternberg, Antje von (Hrsg.), Demokratie und künstliche Intelligenz, Demokratische Meinungsbildung und künstliche Intelligenz, 2019.

Uscinski, Joseph E. / Butler, Ryden W., The Epistemology of Fact Checking, Critical Review 2013, 162–180.

Vargo, Chris J / Guo, Lei / Amazeen, Michelle A, The agenda-setting power of fake news: A big data analysis of the online media landscape from 2014 to 2016, New Media & Society 2018, 2028–2049.

Varol, Onur / Ferrara, Emilio / Davis, Clayton A / Menczer, Filippo / Flammini, Alessandro, Online Human-Bot Interactions: Detection, Estimation, and Characterization, Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017) 2017, 280–289.

Vosoughi, Soroush / Roy, Deb / Aral, Sinan, The spread of true and false news online, Science 2018, 1146–1151.

Walter, Nathan / Cohen, Jonathan / Holbert, R. Lance / Morag, Yasmin, Fact-Checking: A Meta-Analysis of What Works and for Whom, Political Communication 2020, 1–26.

Walter, Nathan / Murphy, Sheila T., How to unring the bell: A meta-analytic approach to correction of misinformation, Communication Monographs 2018, 423–441.

Wank, Rolf, Gewaltenteilung - Theorie und Praxis in der Bundesrepublik, Jura 1991, 622–628.

Wardle, Claire, The Need for Smarter Definitions and Practical, Timely Empirical Research on Information Disorder, Digital Journalism 2018, 951–963.

Wardle, Claire / Derakhshan, Hossein, Information Disorder: Toward an interdisciplinary framework for research and policy making, 2017.

Wilson, Richard Ashby / Land, Molly K, Hate Speech on Social Media: Towards a Context-Specific Content Moderation Policy, Connecticut Law Review 2020, 1–47.

Wolfe, Christopher R. / Britt, M. Anne, The locus of the myside bias in written argumentation, Thinking & Reasoning 2008, 1–27.

Zerback, Thomas / Töpfl, Florian / Knöpfle, Maria, The disconcerting potential of online disinformation: Persuasive effects of astroturfing comments and three strategies for inoculation against them, New Media & Society 2021, 1080–1098.

Zhang, Jerry / Carpenter, Darrell / Ko, Myung, Online astroturfing: A theoretical perspective, Proceedings of the Nineteenth Americas Conference on Information Systems 2013, 1–7.

Zimmermann, Fabian / Kohring, Matthias, „Fake News" als aktuelle Desinformation. Systematische Bestimmung eines heterogenen Begriffs, M&K 2018, 526–541.

Zipperle, Florian, Überblick über Botnetz-Erkennungsmethoden, Seminar Network Architectures and Services, 2014, 17–23.

Zollo, Fabiana, Dealing with digital misinformation: a polarised context of narratives and tribes, EFS2 2019, 1–15.

Zuiderveen Borgesius, Frederik J. / Möller, Judith / Kruikemeier, Sanne / Ó Fathaigh, Ronan / Irion, Kristina / Dobber, Tom / Bodo, Balazs / De Vreese, Claes, Online Political Microtargeting: Promises and Threats for Democracy, ULR 2018, 82–96.

IMPRINT

Publisher:
Media Authority of North Rhine-Westphalia
Zollhof 2
D-40221 Düsseldorf

info@medienanstalt-nrw.de
www.medienanstalt-nrw.de

Project Manager:
Sabrina Nennstiel
Head of Unit Communications, Media Authority of North Rhine-Westphalia
Dr. Meike Isenberg
Head of Unit Research, Media Authority of North Rhine-Westphalia

Coordination and Content Support:
Dr. Meike Isenberg, Thomas Wierny

Authors:
Dr. Stephan Dreyer, Elena Stanciu, Keno Potthast, Prof. Dr. Wolfgang Schulz
(Leibniz-Institut für Medienforschung | Hans-Bredow-Institut)

Design:
Merten Durth (disegno kommunikation GbR)