The ethics of fighting disinformation

Marko Kovic* Adrian Rauchfleisch†

March 2023

Abstract

The global rise of digital disinformation has prompted academia, policymakers and civil society to develop and deploy interventions against disinformation. However, such interventions can also cause damage by restricting the very principles of deliberative democracy they seek to protect. This benefit-vs.-harm conundrum poses an important ethical challenge: How much intervention harm is too much? In this paper, we develop an analytical framework for evaluating the ethical status of disinformation interventions. We proceed in four steps. First, we propose a taxonomy of disinformation interventions. Second, we discuss the available evidence for the effectiveness of the various intervention types. Third, we evaluate the potential damage of disinformation interventions from a consequentialist perspective. Fourth, we combine our findings in a framework that weighs effectiveness against risk. We argue that the group of high net-benefit interventions is ethically unobjectionable, whereas the group of high-impact high-damage interventions, which can be thought of as deliberative weapons of mass destruction, should be used with great restraint. The main benefit of our proposed framework is that it is not a static one-time assessment but rather a generalized and dynamic tool that can be updated with future research: As the evidence on intervention impact grows and becomes more precise, so do the ethical assessments generated with the framework.

*marko@kovic.ch
†adrian.rauchfleisch@gmail.com

Contents

List of Figures

List of Tables

1 Zusammenfassung

Desinformation ist Falschinformation, die bewusst und absichtlich gestreut wird, um damit Ziele zu erreichen. Der Einsatz von Desinformation in der Politik ist kein neues Phänomen. Im heutigen von Online-Vernetzung und Social Media-Plattformen geprägten Kommunikationsumfeld ist die Gefahr von Desinformation aber deutlich gestiegen: Digitale Desinformation kann mit relativ geringen Kosten an sehr grosse Publika herangetragen werden.

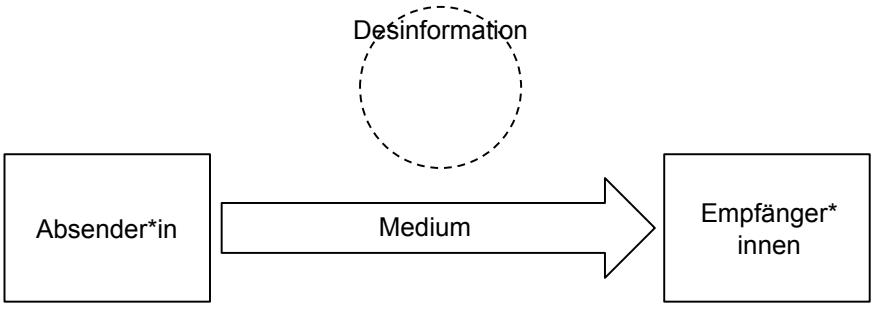
Angesichts der zunehmenden Bedrohung durch Desinformation steigen auch die wissenschaftlichen, zivilgesellschaftlichen und politischen Bemühungen, Interventionen einzusetzen, um Desinformation und den von ihr angerichteten Schaden zu reduzieren. In der Diskussion rund um Interventionen gegen Desinformation hat ein Aspekt bisher zu wenig Beachtung erhalten: Die Frage, ob die eingesetzten Interventionen selber auch Schaden anrichten. Nämlich dann, wenn die Interventionen jene demokratischen Werte, die sie eigentlich schützen sollen – freie Rede und deliberativen Pluralismus – selber beschneiden.

Die vorliegende Studie widmet sich dieser Fragestellung. Die Arbeit ist in vier Schritten aufgebaut: Zunächst erstellen wir eine *Taxonomie der Interventionen*, prüfen danach die *Wirksamkeit dieser Interventionen*, evaluieren anschliessend den *Schaden der Interventionen* und fügen die Ergebnisse in einem *ethischen Framework* zusammen.

1.1 Taxonomie der Interventionen

Ausgangslage für unsere Taxonomie der Desinformations-Interventionen ist das in Abbildung 1 abgebildete minimale Modell der Desinformations-Verbreitung.

Abbildung 1: *Modell der Desinformations-Verbreitung.*



Ein*e Absender*in von Desinformation verbreitet ihre Desinformation über ein bestimmtes Medium an ein Publikum. Aus diesem Modell bilden wir drei Obertypen unserer Taxonomie, die angeben, wo Interventionen gegen

Desinformation ansetzen können: Bei den *Absender*innen*, die Desinformation verbreiten; bei dem *Inhalt* der Desinformation, der über ein Medium verbreitet wird; sowie bei den *Empfänger*innen* der Desinformation.

Innerhalb dieser drei Gruppen verorten wir dreizehn Typen von Desinformations-Interventionen. Diese sind in Tabelle 1 zusammengefasst.

Tabelle 1: *Taxonomie von Desinformations-Interventionen.*

Ort der Intervention	Intervention
Absender*in	Blockieren
	Deplatforming
	Verifikation forcieren
	Sichtbarkeit von Quellen reduzieren
	Quellen kennzeichnen
Inhalt	Inhalte löschen
	Sichtbarkeit von Inhalten reduzieren
	Inhalte kennzeichnen
Empfänger*innen	Feuer mit Feuer
	Feuer mit Wasser
	Nudging
	Fact-checking
	Prebunking

Auf der Ebene der Absender*innen gibt es fünf Interventionstypen. *Blockieren* bedeutet, einem Akteur komplett den Zugang zu einem diskursiven Raum (z.B. einem Land) zu verwehren. *Deplatforming* bedeutet, einen Akteur nur selektiv auf einzelnen Plattformen zu sperren. *Verifikation zu forcieren* bedeutet, auf Social Media-Plattformen keine anonymen Konten zuzulassen. *Sichtbarkeit von Quellen zu reduzieren* bedeutet, im Sinne des "Shadowbanning" Akteuren Zugang zu Social Media-Plattformen zu gewähren, ihre Inhalte aber kategorisch algorithmisch zu unterdrücken und damit ihre Reichweite einzuschränken. *Quellen kennzeichnen* bedeutet, auf Social Media anzugeben, dass beispielsweise eine staatlich kontrollierte Medienagentur ebendies ist.

Auf der Ebene des Inhaltes gibt es drei Interventionstypen. *Inhalte löschen* bedeutet schlicht, einzelne Inhalte z.B. auf Social Media gezielt zu entfernen. *Sichtbarkeit von Inhalten reduzieren* bedeutet, einzelne Inhalte auf Social Media algorithmisch zu unterdrücken und damit ihre Reichweite zu senken. *Inhalte kennzeichnen* bedeutet, bei einzelnen Inhalten auf Social Media Zusatzinformationen anzubringen, wie zum Beispiel Links zu weiterführenden Texten zu einem kontroversen Thema.

Auf der Ebene der Empfänger*innen schliesslich gibt es fünf Interventionstypen. *Feuer mit Feuer bekämpfen* bedeutet, als Reaktion auf Desinformation Desinformation zu verbreiten. *Feuer mit Wasser bekämpfen* bedeutet, dafür zu sorgen, dass im Imformationsumfeld viele hochwertige Quellen und Inhalte z.B. journalistischer Natur gegeben sind, die indirekt gegen Desinformation schützen. *Nudging* bedeutet, irrationale Denkmuster auszunutzen, um Menschen zu vorsichtigerem Umgang mit Desinformation zu bewegen (z.B. in Form von Appellen, zu überlegen, ob geteilte Inhalte zuverlässig sind). *Fact-checking* bedeutet, Desinformation inhaltlich zu prüfen und durch korrekte Informationen zu entkräften. *Prebunking* bedeutet, Menschen präventiv mit einer sprichwörtlichen kognitiven Impfung gegen Desinformation resilient zu machen, entweder über präventive Faktenchecks oder mittels Aufklärung über Logikfehler u.ä. in Desinformations-Narrativen.

Diese dreizehn Interventionstypen sind grundsätzlich universal und können unabhängig von den konkret relevanten Kommunikationsvektoren eingesetzt werden. In der aktuellen Debatte sind aber vor allem digitale Kanäle wie Social Media Fokus des Interesses: So, wie digitale Desinformation grossen Schaden anrichten kann, können digitale Interventionen gegen Desinformation potenziell grosse Wirkung entfalten.

1.2 Wirksamkeit der Interventionen

Wir haben die Effektivität oder Wirksamkeit der Interventionen anhand der verfügbaren Evidenz eigeschätzt, und zwar auf zwei Ebenen: Jener der *engen* und jener der *breiten* Wirksamkeit. Enge Wirksamkeit ist Wirksamkeit im Rahmen des spezifischen Einsatzbereiches einer Intervention. Studien beispielsweise, die experimentell in einem Labor-Setting messen, was für einen Effekt eine Nudging-Intervention hat, messen Wirksamkeit im engeren Sinn. Breite Wirksamkeit betrifft die Frage, ob solche Nudging-Interventionen im realen Praxiseinsatz einen nennenswerten Beitrag gegen Desinformation leisten.

Aus diesen zwei Einschätzungen ergibt sich pro Interventionstypus eine Wirksamkeits-Kennzahl auf einer Skala von 0 (keine Wirksamkeit) bis 4 (hohe Wirksamkeit). Die Wirksamkeiten der Interventionen sind in Tabelle 2 zusammengefasst. Die detaillierte Begründung für die Einschätzung der Wirksamkeiten findet sich in Abschnitt 4.

1.3 Schaden der Interventionen

Wir schätzen den potenziellen Schaden der Interventionen aus der Perspektive deliberativer Demokratietheorie ein. Interventionen sind dann schädlich, wenn

Tabelle 2: *Wirksamkeit von Desinformations-Interventionen.*

Ort der Intervention	Intervention	Wirksamkeit
Absender*in	Blockieren	4
	Deplatforming	3
	Verifikation forcieren	3.5
	Sichtbarkeit von Quellen reduzieren	2.5
	Quellen kennzeichnen	2
Inhalt	Inhalte löschen	1.5
	Sichtbarkeit von Inhalten reduzieren	1.5
	Inhalte kennzeichnen	0.5
Empfänger*innen	Feuer mit Feuer	4
	Feuer mit Wasser	0.5
	Nudging	1.5
	Fact-checking	1.5
	Prebunking	1.5

sie Akteuren, die nicht Desinformation verbreiten, die Teilnahme am Diskurs erschweren oder verunmöglichen; wenn sie latente oder direkte Zwänge schaffen; und, wenn sie selber trügerisch sind. Das Schadensrisiko ist in Tabelle 3 zusammegefasst; die detaillierte Begründung für die Werte findet sich in Abschnitt 5.

Tabelle 3: *Schadensrisiko der Desinformations-Interventionen.*

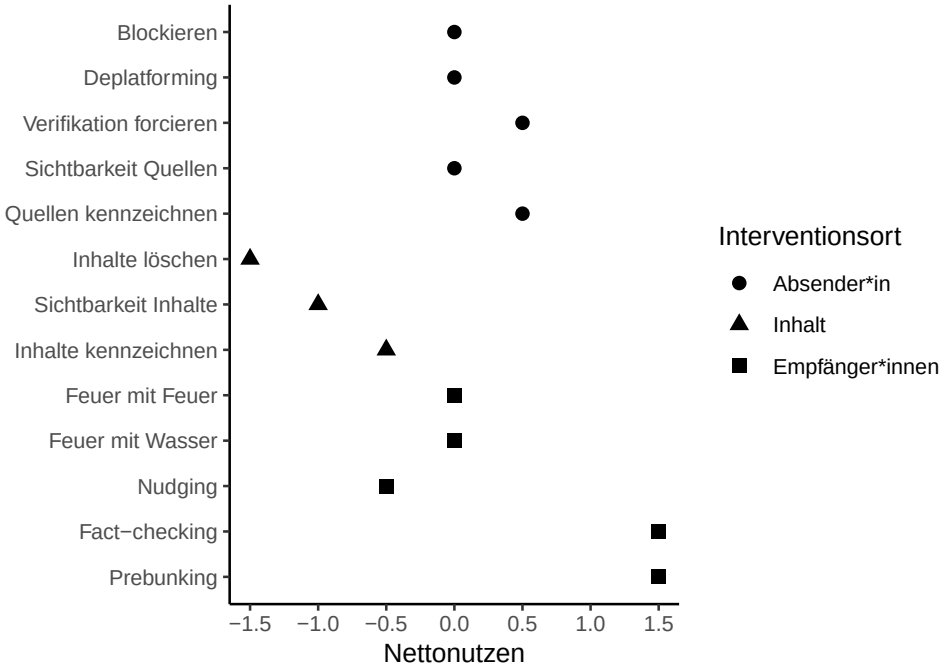
Ort der Intervention	Intervention	Schadensrisiko
Absender*in	Blockieren	4
	Deplatforming	3
	Verifikation forcieren	3
	Sichtbarkeit von Quellen reduzieren	2.5
	Quellen kennzeichnen	1.5
Inhalt	Inhalte löschen	3
	Sichtbarkeit von Inhalten reduzieren	2.5
	Inhalte kennzeichnen	1
Empfänger*innen	Feuer mit Feuer	4
	Feuer mit Wasser	0.5
	Nudging	2
	Fact-checking	0
	Prebunking	0

1.4 Ethisches Framework

Die ethische Bewertung der Interventionen findet in zwei Schritten statt. Zunächst berechnen wir den Nettonutzen der Interventionen, indem wir die jeweiligen Werte für das Schadensrisiko von den Werten für Wirksamkeit abziehen. In einem zweiten Schritt stellen wir die Dimensionen der Wirksamkeit und des Schadensrisikos expliziter zueinander in Beziehung, um die ethische Einordnung insbesondere bei Interventionen mit hoher Wirksamkeit und hohem Schaden zu präzisieren.

Der Nettonutzen der Interventionen ist in Abbildung 2 abgebildet. Die Interventionen mit dem höchsten Nettonutzen sind Fact-checking und Prebunking. Den tiefsten Netto-Nutzen weisen das Löschen von Inhalten und die Reduktion der Sichtbarkeit von Inhalten auf.

Abbildung 2: *Nettonutzen der Interventionen.*

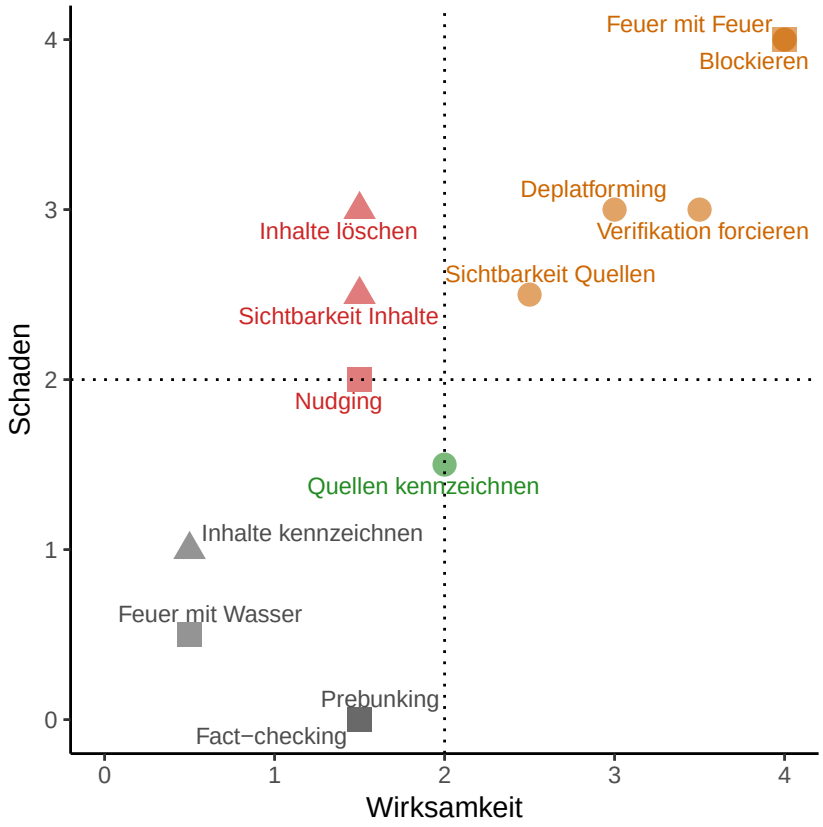


Die detailliertere Gegenüberstellung von Wirksamkeit und Schaden der Interventionen ist in Abbildung 3 abgebildet. Die Darstellung ist in vier Quadranten unterteilt. Der Quadrant unten links weist geringe Wirksamkeit und geringen Schaden auf. Die Interventionen in diesem Quadranten sind tendenziell unbedenklich, aber hinsichtlich des Nutzens auch nicht besonders attraktiv. Eine Ausnahme bilden Fact-checking und Prebunking, die zwar

geringe Wirksamkeit aufweisen, dafür aber keinen Schaden verursachen (Aus diesem Grund haben sie in Abbildung 2 den höchsten Nettonutzen.).

Der Quadrant unten rechts ist der Quadrant mit hoher Wirksamkeit und tiefem Schaden. Interventionen in diesem Quadranten sind aus ethischer Sicht tendenziell am wünschenswertesten. In unserer Analyse fällt aber nur eine Intervention knapp in diesen Quadranten: Das Kennzeichnen von Quellen.

Abbildung 3: *Wirksamkeit und Schaden der Interventionen.*



*Bemerkung: Kreise sind Absender*in-fokussierte Interventionen; Dreiecke sind Inhalts-fokussierte Interventionen; Quadrate sind Empfänger*innen-fokussierte Interventionen. Die Farben geben die Position im jeweiligen Quadranten an. Die Datenpunkte für Prebunking und Fact-checking (Quadrant unten links) sowie Blockieren und Feuer mit Feuer (Quadrant oben rechts) überlappen sich.*

Der Quadrant oben links enthält Interventionen, die eher geringe Wirksamkeit, dafür aber hohes Schadensrisiko haben. Interventionen in diesem

Quadranten sind aufgrund des überproportional grossen Schadenpotenzials zu meiden.

Der Quadrant oben rechts enthält Interventionen, die sowohl hohe Wirksamkeit als auch hohes Schadensrisiko haben. Diese Interventionen sind heikel, weil sie trotz des hohen Schadensrisikos aufgrund ihrer hohen Wirksamkeit attraktiv sein können. Interventionen in diesem Quadranten ähneln vom Impact her Massenvernichtungswaffen: Sie sind sehr wirksam, richten gleichzeitig aber grossen Schaden an. Die Nutzung dieser Interventionen bedarf einer expliziten Begründung, die auch demonstriert, dass die Intervention, welche eingesetzt werden soll, mit hoher Präzision eingesetzt werden kann, damit deliberativer Kollateralschaden vermieden werden kann.

1.5 Policy-Empfehlungen

Aus unserer Analyse ergeben sich drei Regeln, welche in der Praxis beachtet werden sollten.

Erstens muss bei der Erwägung von Desinformations-Interventionen das *Vorsichtsprinzip* zum Einsatz kommen. Desinformation ist ein Untertypus von Falschinformation, aber Falschinformation ist nicht immer Desinformation. Aus demokratietheoretischer Sicht sind Massnahmen gegen gezielte und manipulative Desinformation legitim, Massnahmen gegen aufrichtig geglaubte Falschinformation hingegen nur bedingt: Menschen haben das deliberative Recht, irrational zu sein und ihre entsprechende Meinung kundzutun. In der Praxis ist es schwierig, die Motivsturktur hinter öffentlich kommunizierter Falschinformation festzustellen, also zu klären, ob ein Akteur absichtlich Desinformation verbreitet oder nicht. Im Zweifelsfall muss darum das Vorsichtsprinzip gelten, bei dem davon ausgegangen wird, dass es sich um aufrichtige Falschinformation und nicht um absichtliche Desinformation handelt. Nur so lassen sich falsch-positive Einschränkungen deliberativer Freiheiten vermeiden.

Zweitens sind zunächst jene Interventionen gegen Desinformation ethisch legitim, welche einen positiven *Nettonutzen* aufweisen. Im Rahmen unserer in Abbildung 2 zusammengefassten Analyse haben vor allem Fact-checking und Prebunking einen deutlich positiven Nettonutzen. In weiteren Studien und mit neuer Evidenz wird sich die Einschätzung des Nettonutzens womöglich verschieben. Was in praktischer Hinsicht zählt, ist das analytische Prinzip des Nettonutzens an sich: Ein positiver Nettonutzen ist ein erster Indikator für ethische Akzeptabilität.

Drittens muss jenseits des eindimensionalen Nettonutzens auch das zweidimensionale Verhältnis von *Wirksamkeit und Schadensrisiko* mitberücksichtigt werden. Besonders die Gruppe von Interventionen mit hoher Wirksamkeit

und hohem Schadensrisiko ist in diesem Kontext von Bedeutung. Der Nettonutzen bei diesen Interventionen mag zwar bei rund Null liegen, aber sie haben einen doppelt hohen Impact. Bei der Evaluation von Interventionen in dieser Gruppe plädieren wir für eine Asymmetrie zwischen Wirksamkeit und Schadensrisiko: Wenn beide Werte hoch sind, ist der Schaden höher zu gewichten als die Wirksamkeit. Der Schaden ist ethisch schlechter als die Wirksamkeit gut ist. Der Einsatz solcher Interventionen ist darum nur legitim, wenn explizit begründet werden kann, dass ein geplanter Einsatz mit hoher Präzision stattfinden kann, sodass das Schadensrisiko – deliberative Falsch-Positive – signifikant gesenkt wird.

2 Introduction: Fighting disinformation, but at what cost?

On February 24 2022, Russia invaded Ukraine in a major escalation of the Russo-Ukrainian war that started in 2014. On March 2 2022, as a reaction to the invasion, the European Union banned the Russian state-sponsored media outlets RT (formerly "Russia Today") and Sputnik within the borders of the European Union. The goal of the ban was to stem the flow of Russian war-related disinformation into EU countries. RT and Sputnik are important disinformation vectors in the Russian "firehose of falsehoods" propaganda strategy [1], and their often conspiratorial messaging [2] appeals to and to some degree radicalizes people who distrust traditional media and institutions, and who suffer, relatively speaking, conditions of material deprivation [3, 4].

Given the dramatic events that precipitated the EU ban of RT and Sputnik, the decision seems intuitively morally just. The effects of the ban, however, are both practically and ethically opaque. Practically, it is unclear how effective the ban really was, given that the Kremlin rapidly employed tactics to circumvent the ban [5]. While it seems plausible to assume that Russian disinformation campaigns were at least temporarily disrupted by the ban, EU officials have no metrics to assess its impact.

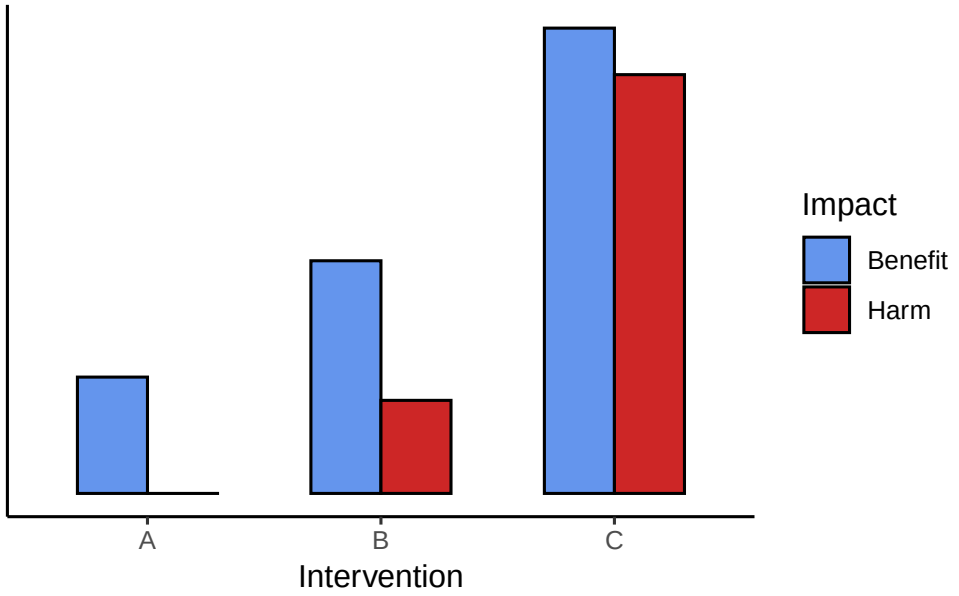
Ethically, the ban of RT and Sputnik, while legally sound, is questionable from a normative democratic perspective [6]. A blanket ban of Russian state-sponsored media outlets collides with the democratic ideals of freedom of speech and pluralism. The European Union justified its ban by pointing to the potential damage Russian disinformation does, but it failed to weigh that damage against the potential damage the ban itself does in terms of limiting democratic freedoms.

The ban of RT and Sputnik in the wake of the 2022 invasion of Ukraine is an example of a broader dilemma in the fight against political disinformation. Disinformation is misinformation that is actively created and disseminated in order to achieve some goal [7, 8]. As such, it is inherently anti-democratic because it undermines, at the very least, the process of rational and genuine political deliberation [9, 10]. Stopping or reducing this type of anti-democratic sabotage is essentially self-evidently important. At the same time, however, interventions against disinformation can potentially, and somewhat paradoxically, harm the very principles and freedoms they aim to protect. Disinformation works by exploiting deliberative pluralism and freedom of speech, both of which are cornerstones of democratic societies. If interventions tackle disinformation by curtailing deliberative pluralism and freedom of speech, those interventions are, though well-intended and potentially effective,

ethically questionable.

To illustrate this point, consider the simple generalization of this problem in the thought experiment presented in Figure 4.

Figure 4: *Disinformation intervention impact thought experiment.*



Depicted is the impact of three fictional disinformation interventions. If we only evaluate the three interventions based on the benefit they confer, intervention C is clearly superior to both A and B, and B is clearly superior to A. If we, however, also take the harm the interventions themselves do into account, the evaluation is less obvious. Is A the preferred intervention because it does no harm? Or is B the preferred intervention because it does more good than A and has a favorable benefit-to-harm ratio (in that its net positive impact is higher than A)? Or is C still the superior intervention because it has the highest total benefit and it does, overall, more good than harm?

This simple thought experiment demonstrates that the moral evaluation of disinformation interventions changes and becomes more complex when the potential (extent of the) harm of interventions is taken into account. The importance of finding a justifiable ethical balance between benefits and risks of disinformation interventions has been stressed before [11], but the precise nature of that balance has so far not yet been explicitly addressed.

2.1 A consequentialist approach

While there have been several critical reviews on misinformation and disinformation interventions and their effectiveness in recent years [12, 13, 14], the debate on ethics of disinformation interventions is still sparse, but it is growing. In their study on the effect of labeling sources on Twitter, Aguerri et al. [15] note that de-amplifying content consititues a restriction on freedom of speech, which, the authors argue, needs to be problematicized. In an experimental study on content moderation preferences of regular social media users, Kozyreva et al. [16] find that study participants' moral intuition favors removing misinformation rather than maximizing freedom of expression.

The most direct contribution so far is by Bjola [17]. Bjola argues that the moral acceptability of disinformation interventions is contingent on the concept or moral authority: A state or organization needs to make the case that it has been harmed by disinformation; that it has normative standing to engage in counter-interventions (in that it has accountability, integrity and effectiveness); and that it does so in a proportionate and responsible manner.

Bjolas' argument is a deontological [18] one. He specifies a set of conditions that have to be met in order for a disinformation intervention to be just and called for. Our view is different: We argue that the moral standing of a disinformation intervention is determined by the outcomes or consequences it produces. Our moral outlook is therefore consequentialist [19] in nature.

In this paper, we analyze two types of consequences of disinformation interventions: Their effectiveness and their potential harm. We then analyze these two dimensions in our proposed ethical framework, which consists of two steps: A symmetrical and an asymmetrical analysis. In the symmetrical analysis, we regard the benefit of disinformation interventions to be as good as their harm is bad. What determines the ethical status of an intervention is its net benefit. This analysis is similar to the concept of total utilitarianism within utilitarian moral philosophy, whereby the focus is on the total amount of welfare or value that results after suffering and happiness, which are weighed equally, are added up. This first analytical step of symmetrically calculating net benefits shows which disinformation interventions are safe to use given their positive net benefits.

In the asymmetrical analysis, we regard effectiveness and harm to be unequal in their value and disvalue. In this view, harm is worse than the good done by effectiveness, which shifts the overall evaluation. This analysis is analogous to the suffering-focused argument within utilitarianism that posits an asymmetry between suffering and happiness [20, 13-110]. We apply this asymmetrical analysis for interventions that have high effectiveness as well as high harm potential. These interventions might seem attractive given their

effectiveness and seemingly acceptable benefit-harm ratios, but the potentially large amount of harm they do means that, as we argue, their use should be limited to cases in which they are explicitly and broadly justified and their damage is demonstrably and significantly minimized.

The symmetrical and asymmetrical ethical analyses are complementary. The symmetrical view is a starting point for revealing unobjectionable interventions with positive net benefits, and the asymmetrical view is a cautionary expansion of the symmetrical analysis that reveals interventions of special concern. Together, these two evaluations form our proposed ethical framework.

That framework, which we hope will further and make more precise the debate on the ethics of disinformation interventions, is the primary purpose of this paper. Our ratings for disinformation effectiveness are a means to this end. They should be understood as preliminary assessments given the current state of evidence and plausibility. Those assessments can and should be updated in the future.

2.2 Structure of this paper

We proceed in four steps. First, we give an overview of disinformation intervention types in section 3. Second, we assess the effectiveness of the intervention types in section 4. Third, we evaluate the harms of the intervention types in section 5. Finally, we combine our analyses of the benefits (effectiveness) and harms of disinformation interventions into a framework in section 6.

3 A taxonomy of interventions

Disinformation is not a new phenomenon [21]. In recent years, however, disinformation has become a more urgent concern both in academic and in public policy discourse, primarily due to the role of social media. Social media platforms allow malicious actors to easily spread disinformation to potentially large and geographically distant audiences [22]. The true extent of disinformation on social media is unknown because disinformation is often deceptive in nature, at best detectable as misinformation of unknown origins. There is some indication that disinformation is used in and by dozens of countries [23]. Digital disinformation, it seems safe to say, is a tool wielded widely and frequently, and the global dynamics of misinformation and disinformation are detrimental to democracy [24].

Given this new sense of urgency, many interventions against disinformation have been tested and implemented in the recent past. The Consortium for

Elections and Political Process Strengthening lists 282 interventions in over 80 countries in its "Countering Disinformation Guide" [25], and, as of 2021, there are at least 100 laws in over 70 countries aimed at curbing misinformation and disinformation [26][1]. The European Union's East StratCom Task Force, established in 2015 as a reaction to the 2014 Russo-Ukrainian war, is a supranational anti-disinformation effort tasked with promoting European values and cataloguing Kremlin-backed disinformation through various means [27]. Given this multitude of real-world attempts at tackling disinformation, creating a general taxonomy of disinformation interventions is challenging.

Existing taxonomies in recent studies focus on disinformation interventions on social media platforms. Alemanno [28] identified three broad intervention categories: Governmental surveillance of disinformation on social media; content liability for social media companies; increasing the volume of accurate information on social media in order to reduce the salience of disinformation. In a survey of real-world actions taken by social media platforms, Yadav [29] identified nine types of interventions: Redirections to additional information; labeling content with additional information; labeling content as mis- or disinformation; reducing options to share content; increasing disinformation literacy; disclosing paid advertisements; adding content reporting options; content and account moderation (including deplatforming actions); increasing security or verification requirements. Kozyreva et al. [12] have created a toolbox in which they identify ten interventions against online misinformation and manipulation: Accuracy prompts, debunking, friction, inoculation, lateral reading, media literacy tips, rebuttals of science denialism, self-reflection tools, social norms and warning and fact-checking labels.

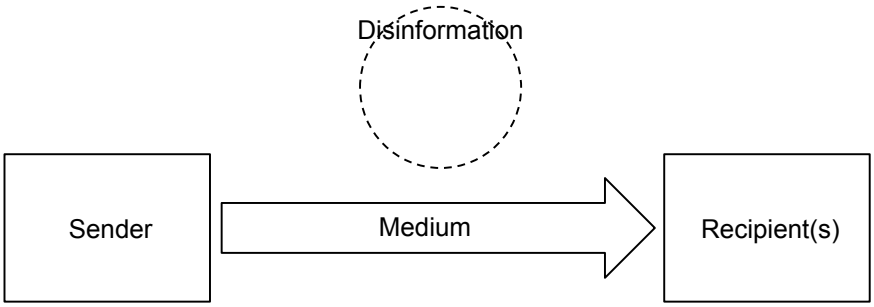
For the purpose of this paper, we propose a taxonomy that includes disinformation interventions on social media platforms but is conceptually not limited to social media. Social media is certainly an important and possibly even the central venue of contemporary disinformation attacks, given that it allows disinformation agents to reach potentially large audiences at relatively low cost [30]. But disinformation can be transmitted on the internet through vectors other than social media platforms, and it can be transmitted through "traditional" broadcast and print media. A generalized taxonomy of disinformation interventions should take this multi-modal nature of disinformation into account.

The starting point for our taxonomy is a minimal model of disinformation propagation based on Lasswell's communication model [31]. The model, as depicted in Figure 5, consists of four basic elements: A disinformation sender

[1]Many of those laws, however, have been enacted in authoritarian countries where the supposed fight against mis- and disinformation is a pretext for suppressing dissent.

who communicates a piece of disinformation through some medium to one or typically multiple recipients who are to be influenced by the disinformation. This model is not a complete account of disinformation propagation, which would have to include additional elements such as social diffusion [32]. The simplified model we apply is a conceptual abstraction of the basic building blocks of disinformation propagation that serves as the foundation of our taxonomy of intervention types.

Figure 5: *Basic model of disinformation propagation.*



The three basic building blocks of this model of disinformation propagation are the three basic locations or points at which interventions against disinformation can be applied. The *sender* (or communicator), the *content* sent through some medium, and the *recipient* form the three overarching categories of our taxonomy.

In the disinformation *sender* category, there are five types of interventions:

- *Blocking*: Blocking a communicator means preventing or stopping them from communicating at all within a certain jurisdiction. An example of this strategy is the blocking of the Russian state-sponsored media outlets RT and Sputnik in the countries of the European Union.

- *Deplatforming*: Deplatforming is partial or selective blocking whereby a sender of disinformation cannot use a certain or multiple communication vectors any longer but retains some legal access to the targeted public through other communication vectors.

- *Forcing verification*: Forcing verification means making access to platforms conditional on providing evidence of a person's or an organisation's identity in order to, for example, stop a disinformation sender from using fake sockpuppet accounts on social media.

- *Reducing source visibility*: Reducing a disinformation sender's visibility means allowing them access to communication vectors, but making the

(dis-)information they spread through these vectors categorically harder to access than information spread by regular, non-sanctioned senders.

- *Labeling sources*: Labeling sources means adding prominently visible information about the nature of a source to that source. For example, Twitter labels some news outlets controlled by governments as "state-affiliated media".

The interventions in the disinformation *content* category do not affect the sender of disinformation directly or categorically but are instead selectively applied to their disinformation messages within the communication vectors they use. There are three types of intervention in this category:

- *Deleting content*: Deleting content means selectively deleting singular pieces of (dis-)information that an actor is spreading through some communication vector or vectors.

- *Reducing content visibility*: Reducing content visibility means that (dis-)information disseminated through some communication vector or vectors is not deleted but instead made selectively harder to access than other kinds of information. In contrast to the visibility reduction intervention at the sender level, the intervention at the content level only selectively targets pieces of (dis-)information rather than categorically all (dis-)information spread by a sender.

- *Labeling content*: Labeling content means appending additional information to published (dis-)information in order to reduce interactions with the specific (dis-)information in question.

Interventions in the *recipient* category are meant to positively affect or protect the targets of disinformation who are at risk of being manipulated by it. The main target is typically the general public that consists of individual people, but publicly disseminated disinformation can also target organizations such as news media, private businesses or political actors. There are four types of interventions in this category:

- *Fire with fire*: Fighting fire with fire means deploying disinformation that is supposed to counter disinformation spread by another actor. The logic of such an intervention is to expose recipients of disinformation to disinformation of another epistemic and political bent in order to counteract undesired effects. A historical example of this approach are Western propaganda radio stations meant to counteract Soviet propaganda radio [33].

- *Fire with water*: Fighting fire with water means creating an information environment in which disinformation is surrounded by correct information. By increasing the amount of credible or correct information in an information environment, disinformation's relative visibility and psychological impact are reduced, even though the correct information does not directly address the disinformation in question. An example of this approach are links to credible resources that are automatically added on social media when users post about disputed issues [28].

- *Nudging*: Nudging means changing decision-making contexts by exploiting cognitive heuristics and biases in order to induce a desired behavior [34]. In the context of misinformation and disinformation, nudging can take the form of behavioral prompts on social media that affect users' perception of and interaction with content, primarily in terms of sharing behavior, without forcing them to behave any specific way.

- *Fact-checking*: Fact-checking or debunking means presenting corrective information about some piece of misinformation or disinformation that is in circulation. We use the terms debunking and fact-checking as synonyms in this paper. In contrast to the fire with water approach, fact-checking explicitly addresses the claims made by the disinformation in question.

- *Prebunking*: Prebunking is an umbrella term for preventative interventions that provide corrective information before an individual comes into contact with misinformation or disinformation. The goal of this type of intervention is to either debunk specific disinformation claims before they circulate widely in the public discourse (fact-based prebunking), or to provide potential recipients of disinformation with knowledge and cognitive skills that allow them to detect and deal with dubious claims in general (logic-based prebunking). Prebunking interventions are based on the principles of psychological inoculation [35].

In total, our proposed taxonomy of disinformation interventions contains thirteen types of interventions. They are summarized in Table 4. The taxonomy is not an exhaustive list of specific interventions – there can be numerous varieties within each type – but instead a description of the conceptual space of disinformation interventions. Our taxonomy is a list of strategies for tackling disinformation.

Table 4: *Taxonomy of disinformation interventions.*

Intervention locus	Intervention type
Sender	Blocking
	Deplatforming
	Forcing verification
	Reducing source visibility
	Labeling sources
Content	Deleting content
	Reducing content visibility
	Labeling content
Recipient	Fire with fire
	Fire with water
	Nudging
	Fact-checking
	Prebunking

4 Intervention effectiveness

In this section, we evaluate the effectiveness of disinformation intervention types in the form of a narrative review. To that end, we conducted a broad search of the existing literature and analyzed the available evidence for the different types of interventions.

4.1 Defining effectiveness

In the context of disinformation interventions, effectiveness means at least two things. On an individual level, the effectiveness of interventions is a function of how well interventions reduce what we can call epistemic damage: The better an intervention reduces the probability of an individual believing a disinformation claim, the more effective it is[2]. On a societal level, disinformation interventions are effective if they mitigate the democratic damage done by the collective epistemic damage of disinformation. In other words, if interventions reduce the risk of a disinformation agent achieving their large-scale goals, the intervention is effective.

A major shortcoming of the existing literature on disinformation and misinformation interventions is that it provides only limited evidence for

[2]Or, if we conceptualize belief in more Bayesian terms: The better an intervention is at preventing an individual's posterior probability of shifting in favor of a disinformation claim, the more effective the intervention is.

individual-level effects and even less evidence for macro-level effects. Not because the research is faulty, but because it is demanding, and there is currently still a lot of uncertainty.

Most empirical research on interventions is performed on individuals in laboratory experimental conditions. It is largely unknown, as Altay [13] notes, whether the observed individual-level experimental efficacy of interventions translates into real-world individual-level effectiveness. Additionally, as Kozyreva et al. [12] point out, most studies are conducted in countries of the Global North and thus potentially lack generalizability because real-world contextual factors are not taken into account. The evidence of macro-level, societal impacts of misinformation and disinformation interventions is, consequently, also limited. A number of studies, for example, track the dissemination dynamics of disinformation on social media platforms over time, but given the complexity of the problem and restrictions such as limited data availability [36], such studies are typically limited in scope.

These limitations, as we discuss below, are present for all intervention types, which means that much of the hoped for benefit of the different interventions is ultimately speculative rather than quantifiably empirical in nature. This is an important aspect that factors into our ethical framework. We address the misinformation vs. disinformation distinction in greater detail in section 5.

Given the limitations of the available evidence as well as the heterogeneity of the types of evidence, we quantify intervention effectiveness on two generic scales: *Narrow effectiveness* and *broad effectiveness*. Narrow effectiveness refers to the effectiveness of an intervention in the narrow context of the intervention. Broad effectiveness, on the other hand, refers to the overall impact of an intervention on the overall problem of disinformation. For example, forcing social media users to verify their identities might be highly effective in reducing the use of fake accounts for spreading disinformation (narrow effectiveness), but it might have limited effectiveness for reducing disinformation in general (broad effectiveness) because malicious agents can use other means of spreading disinformation. We rate both forms of effectiveness on a scale of 0 (no effectiveness) to 4 (high effectiveness) and calculate a final effectiveness score by adding both scores and then dividing by two.

4.2 Sender interventions

4.2.1 Blocking

There is no research on the effectiveness of attempts at complete blocking of disinformation senders. The reason for that is probably that there seems not

much to be studied: If a disinformation sender is completely stopped from disseminating disinformation, logic dictates that the problem is completely solved – the disinformation in question never has the chance to do damage, and the disinformation agent has by definition failed.

But the impact of blocking is probably less clear-cut than that because senders can, at least to some degree, circumvent the blocking efforts on the internet. In the case of RT and Sputnik which were blocked in the wake of the 2022 Russo-Ukrainian war, the Russian government used alternative domain names to circumvent the blocking. Other large-scale blocking efforts, such as China's "Great Firewall", a prominent part of the larger "Golden Shield Project" project, indirectly demonstrate that more resourceful users are able to circumvent blocking efforts on their ends as well [37, 38]. However, the Great Firewall example also demonstrates that even though that specific blocking measure is porous, it is ultimately still highly effective because of the added friction of having to, for example, pay for and employ tools such as virtual private networks in order to circumvent the censorship is a hurdle too high for most people [39, 56-80].

Another potential downside to blocking disinformation senders is that those senders and their disinformation can be, counter-intuitively, amplified rather than silenced for some parts of the audience. For example, the European ban on RT and Sputnik in March 2022 seemingly made those outlets more popular in some conspiracy-minded online communities which interpreted the ban as censorship of the forbidden truth [40, 41]. This type of "Streisand effect" whereby banning information makes it more attractive and prominent has also been observed in the aforementioned case of the Great Firewall in China [39, 50-54].

Overall, there is little direct evidence on the effectiveness of general blocking of disinformation senders. Blocking is not impermeable (senders can still get through, and recipients can still seek the senders out), and it may backfire to some degree by amplifying the blocked disinformation senders as censored truth-speakers. Additionally, blocking only works of there is a clear and known source of disinformation that can be blocked. Disinformation senders often use clandestine methods for spreading disinformation, such as astroturfed fake users [42] that are neither detectable *a priori* nor easily blockable.

Despite these limitations and the lack of direct evidence, however, blocking is in all likelihood a highly effective intervention in situations in which it can be applied. If a sender of disinformation can be clearly identified as in the case of publications or organizations, blocking that sender will almost certainly reduce the reach of the disinformation the sender is disseminating. The broader impact of blocking is probably also high. By blocking known

disinformation senders, potentially large downstream negative effects are prevented or at least significantly reduced. Given these plausible impact of blocking, our effectiveness ratings for blocking interventions are high:

- Narrow effectiveness: 4.
- Broad effectiveness: 4.
- Total score: 4.

4.2.2 Deplatforming

Contrary to general blocking, selective deplatforming in the context of extremism and hate on online platforms has been extensively studied. One important general finding is that deplatforming can resemble a game of Whac-A-Mole. Banning users from individual platforms does affect those users' reach and impact in the short-term, but in the medium-term and from a macro-level perspective, the online misinformation and hate ecology has proven to be resilient [43]. Even though deplatforming can reduce the amount of misinformation and extremist content circulated within a specific platform [44, 45], the deplatformed actors and networks of misinformation are not simply eliminated. They tend to shift to other communication vectors such as "alt-tech" platforms [46] where they can continue spreading misinformation. These migratory shifts happen in waves, triggered by deplatforming prominent misinformation spreaders who bring attention to alternative platforms when they migrate to them [47]. The networks which are temporarily disrupted by deplatforming can quickly recover on alternative platforms [48], and the dynamics of disseminating misinformation and hate speech can increase both in speed and damage [49, 50]. Besides such migratory patterns, so-called "re-platforming" [51] has also been observed in the context of deplatforming, whereby banned users return to a platform with new accounts and continue the spread of misinformation.

At the same time, however, there is evidence that even though the online misinformation ecology is resilient, deplatforming can reduce the overall amount of misinformation in circulation. For example, in a study of YouTube deplatforming of right-wing extremists, migratory patterns to an alternative video platform was observed, but the overall reach of the content was significantly reduced [52]. A historical example of the effectiveness of deplatforming is the presence of members and sympathizers of the Islamist terror organisation ISIS on social media. Through repeated waves of deplatforming, the volume and reach of pro-ISIS propaganda on Twitter was successfully and significantly reduced [53, 54].

In summary, the evidence on the effectiveness of deplatforming is mixed, but deplatforming can have high impact if it is a continued effort across multiple platforms. Networks of misinformation and disinformation are resilient and cannot be easily disrupted through singular instances of deplatforming, but deplatforming as an iterative intervention can successfully reduce the volume and reach of misinformation and disinformation over time.

Overall, we estimate that deplatforming has relatively high narrow effectiveness, given that it confers immediate short-term benefits, and equally high broad effectiveness if it is implemented in the form of longer-term deplatforming activities that have a cumulatively positive effect over time. Our ratings for deplatforming interventions are accordingly high:

- Narrow effectiveness: 3.
- Broad effectiveness: 3.
- Total score: 3.

4.2.3 Forcing verification

User anonymity has long been recognized as a major contributor to anti-social, disinhibited behavior on the internet [55]. There is evidence that implementation of identity cues such as user verification decrease the amount of anti-social online behavior [56, 57, 58]. In the context of disinformation, however, there is no direct research on the effectiveness of identity verification: We have found no studies that investigate the effect of user verification policies on disinformation agents' ability to spread disinformation.

From a purely logical point of view, forced verification should be a highly effective intervention in a narrow sense. Disinformation agents frequently use fake social media profiles – both automated bots as well as human-operated sockpuppets – in order to spread disinformation [59]. Removing the option to create non-verified accounts would entirely rob them of this attack vector. Under the condition of forced identity verification, disinformation agents would potentially shift some of their activity to verified accounts operated by people who are they say they are. But it is highly improbable that such a shift would be able to compensate for the lost ability to create and operate numerous fake accounts at once. Today, a single disinformation sender can operate, for example, a thousand fake accounts. Under a forced verification regime, it would take a thousand disinformation senders to operate as many verified accounts. Forced identity verification would make the dissemination of disinformation on social media much more costly.

In summary, even though there is no direct evidence on the impact of forced user identity verification on disinformation senders' ability to operate, we estimate that such an intervention has fairly high narrow effectiveness. All else being equal, implementing an identity verification requirement can significantly hamper malicious agents' disinformation capabilities. This impact would, we estimate, also to some degree translate into broad effectiveness. Even though fake social media accounts are only one way in which malicious agents spread disinformation, the existing literature identifies it as a major one. Reducing these capabilities would therefore have a significant overall impact. We quantify these effectiveness estimates into the following ratings:

- Narrow effectiveness: 4.
- Broad effectiveness: 3.
- Total score: 3.5.

4.2.4 Reducing source visibility

Allowing (disinformation) senders to use communication vectors, but categorically reducing the visibility of their messages, is a tactic that has been implemented in the past, but public information on the modalities of such interventions are limited. In the debate about social media moderation, this type of intervention is often referred to as "shadow banning". The nature or even the existence of this practice was somewhat contested in the past, but there is evidence that shadow banning in a broad sense of reducing source visibility is being employed. Numerous social media users on different platforms have over the years claimed that the content they produce is categorically downranked or hidden without the platform noticing them of why that is the case [60]. Social media platforms themselves have so far been hesitant to communicate whether and how they reduce source visibility. Twitter, for example, has vaguely stated that posts from "bad-faith actors who intend to manipulate or divide the conversation should be ranked lower" [61]. Some limited insights into how this downranking works has been exposed in the December 2022 "Twitter files" [62]. An additional form of indirect evidence for source visibility reduction are attempts at reverse-engineering the dynamics of how content and content creators on social media are seen and interacted with over time [63, 64]. Another example is YouTube's changes in its recommendation algorithm. For example, in February 2019, YouTube almost completely stopped recommending videos from alt-right channels in their video recommendations [65].

Regardless of whether reducing source visibility is already widely employed or not, it is an intervention that could possibly reduce the reach of disinformation agents. If, for example, known disseminators of disinformation are disadvantaged in search results and their content is less likely to be algorithmically presented to users, their activities would have less reach. We therefore estimate the hypothetical narrow effectiveness of reducing disinformation agents' visibility to be non-trivial. In a broader context, however, the potential impact would likely be somewhat lower, since reduced visibility on some platforms would mean only a relatively modest dent in some parts of the often employed firehose strategy of multimodal disinformation dissemination [1]. Given the lack of direct evidence and the limited plausibility of potential effectiveness, our ratings for reducing source visibility interventions are as follows:

- Narrow effectiveness: 3.
- Broad effectiveness: 2.
- Total score: 2.5.

4.2.5 Labeling sources

Informing the public that certain actors are likely to spread disinformation should, in theory, reduce the impact those actors have since the targets of the disinformation become aware of the risks associated with a certain source. Ideally, labeling a source in this manner should trigger something like a persuasion knowledge response [66] whereby people seek to resist the disinformation and thereby develop resilience.

The social media platforms YouTube (since 2018), Twitter, and Facebook (both since 2020) apply information labels for some government-controlled content producers and news outlets. Such labels have been found to be both efficacious in experimental settings [67] as well as effective in real-world settings [68, 69, 15] at reducing user interaction with content disseminated by disinformation agents.

Existing studies suggest high narrow effectiveness of labeling sources. Some degree of caution, however, seems in order because strong effects in early studies in other domains have been known to be the result of the so-called novelty bias whereby effects are initially overestimated [70]. Additionally, Twitter, the most commonly studied platform in the context of disinformation interventions, pairs source labeling with reducing source visibility [71, 15]. The effect of labeling sources is therefore probably lower than reported in real-world studies. It is currently also unclear how well the narrow effects of

source labels translate into broader effects against disinformation. A potential overall outcome is a partial slowdown of the spread of disinformation through labeling sources. Given these effectiveness assessments, our ratings for labeling sources interventions are as follows:

- Narrow effectiveness: 2.
- Broad effectiveness: 2.
- Total score: 2.

4.3 Content interventions

4.3.1 Deleting content

Deleting published content is difficult or even impossible in traditional broadcasting, but it is a basic form of content moderation on social media. Typically, social media platforms delete expressly illegal content or content that violates a platform's terms of service. Deletions are occasionally also applied to misinformation. For example, major social media platforms actively deleted medical misinformation during and related to the Covid-19 pandemic [72].

There is no research on the question of the effectiveness of content deletion. From a purely logical point of view, removing content should have non-trivial impact: Disinformation that is removed cannot do any more harm. However, there are several factors that put the effectiveness of deleting content into question. First, deleting content is typically reactive, meaning that the content in question is circulating for some period of time before being removed – at which point the damage might already be done. Second, content moderation is resource intensive, and systematic monitoring and removal of disinformation is probably simply not feasible [25]. Third, in the context of misinformation, migration patterns of deleted content to alternative platforms that resemble migration patterns of deplatformed users have been observed [73], which indicates that disinformation and misinformation are resilient and might require, similar do deplatforming users and accounts, repeated and sustained action.

Overall, deleting disinformation content probably has some narrow effectiveness, but its broader impact is likely to be very limited. Our effectiveness ratings for deleting content interventions are as follows:

- Narrow effectiveness: 2.
- Broad effectiveness: 1.
- Total score: 1.5.

4.3.2 Reducing content visibility

Social media platforms selectively reduce the visibility of content they deem to be "borderline" [74]. That is typically content that does not violate a platform's terms of use, but that is nonetheless deemed problematic. The reduction in visibility means that content is demoted in rankings and recommendations in order to reduce its reach. For example, various social media platforms have reduced the visibility of misinformation related to the Covid-19 pandemic [75].

There is so far no research on the impact of selective "shadow bans" of content. From a logical point of view, the effectiveness of that intervention should be comparable to outright deleting content. In a narrow sense, reducing the visibility of content is less effective than outright removing it. In a broader sense, merely reducing the visibility of some content (without disclosing it) might be somewhat more effective because it avoids migration to other platforms that might amplify the content's reach. However, as with deleting content, it seems unrealistic that misinformation and disinformation can be categorically made less visible since there is no automated way of clearly identifying misinformation, let alone disinformation where there is the added element of malicious intent. Given the lack of direct evidence and the limited plausibility, our effectiveness ratings for reducing content visibility interventions are low:

- Narrow effectiveness: 2.
- Broad effectiveness: 1.
- Total score: 1.5.

4.3.3 Labeling content

In traditional broadcasting, what you see is generally what you get. Adding information to (dis-)information that is already circulating is difficult. On social media, however, appending additional information to published (dis-)information is conceptually and technically feasible. Evidence on the effectiveness of such interventions is mixed but generally points towards limited to no impact.

In their overview of content labeling interentions, Morrow et al. [76] identify two main subtypes of content labeling: Veracity labeling and contextual labeling. Veracity labeling is additional information appended to content that informs users that some content is disputed or outright false. Contextual labeling is additional information appended to content that is, in contrast to

veracity labeling (which directly warns about the content in question), more general in nature, such as a link to further information about a topic.

The evidence on veracity content labeling points towards little to no effect. Some recent studies report a small to moderate effect in reducing interactions with content [77, 78], whereas others find no such effect [79, 80, 81, 82]. One study even found a backfire effect whereby content labeling led to more rather than fewer interactions with the content [83]. That result, however, might be an outlier; the study in question investigated former president Trump's posts on Twitter, and most social media users or accounts do not receive the attention the former president's did.

The evidence on contextual information content labeling is limited. One study found that adding information about the publishers of content that is being shared generally has no impact on interactions with the content [84]. One study found that contextual information about medical topics on Twitter reduces the dissemination of misinformation [85].

A general downside of content label interventions is the risk of the implied truth effect [86]. Labeling some content can create the perception that content without labels is credible or true. Given the fact that content labeling is always selective, the implied truth effect could lead to an indirect amplification of misinformation and disinformation that happens not to be labeled.

Overall, the available evidence points to at best low narrow effectiveness and low to no broad effectiveness of content labeling interventions. This is reflected in our effectiveness ratings for this type of intervention:

- Narrow effectiveness: 1.
- Broad effectiveness: 0.
- Total score: 0.5.

4.4 Recipient interventions

4.4.1 Fire with fire

Fighting disinformation with disinformation is a tactic employed at least since the Cold War. One notable formalization of this intervention type in recent years has been the passing of the 2017 National Defense Authorization Act in the United States. The law contains the "Countering Foreign Propaganda and Disinformation Act" provision which specifies some structural foundations of public US counter-propaganda operations in other countries [87].

Even though the fighting fire with fire approach is common, the question of effectiveness has so far received little scholarly attention. One study on Radio

Free Europe and Radio Liberty finds qualitative evidence that these "influence operations" had reach and impact, not least because Communist and post-Communist political elites testify to that fact [88]. One phenomenon related to the fighting fire with fire approach are propaganda and disinformation interventions of the Central Intelligence Agency and the Department of Defense in Hollywood productions. Research on that issue has noted that the sheer volume and reach of such disinformation is bound to have an impact on narratives about intelligence agencies and the military [89, 90] (though it is debatable whether this specific intervention is really a reaction to disinformation or simply active disinformation in its own right).

Indirect evidence for the effectiveness of fighting fire with fire interventions is the impact of the fire that is supposed to be fought this way. It is fairly obvious that misinformation and disinformation can do epistemic damage [91] and that epistemic damage can translate into detrimental behavior. Misinformation related to the Covid-19 pandemic, for example, resulted in riskier individual health behaviors, worse individual and community health outcomes, worse psychological wellbeing, loss of trust in political institutions, and even violence [92, 93, 94]. It stands to reason that, given the damage misinformation and disinformation can do, disinformation disseminated as a reaction to existing disinformation should also have similar levels of impact. We therefore give this intervention type a high effectiveness rating:

- Narrow effectiveness: 4.
- Broad effectiveness: 4.
- Total score: 4.

4.4.2 Fire with water

There is no direct evidence on the effectiveness of the fighting fire with water approach whereby the impact of disinformation is dampened by increasing the amount of correct information that is available. In a very general sense, increasing the amount of correct information within an information environment should, all else being equal, reduce the prominence of disinformation within the information environment. The logic of this approach mirrors the (unfortunately named) concept of "crippled epistemologies" [95]. Members of an information environment place belief in the information which is available to them. The more dominant misinformation or disinformation is, the more epistemic damage it is bound to produce. The more correct information is available, the less damage the mis- or disinformation will do.

This is a simple enough principle, but it is unclear how it can be translated into effective interventions. For example, one study has found a positive effect when users who are seeing misinformation are offered an additional selection of articles created by reputable journalistic sources [96]. But it is not clear how such a potentially narrow intervention could be translated into a more broadly effective one. Generally speaking, fire is being fought with water when there is a functioning system of independent and critical journalistic media outlets that offer high-quality, evidence-based reporting as credible alternatives to misinformation and disinformation. There is evidence that such a robust state of public discourse can be achieved by publicly funding independent media organizations that are neither politically controlled nor subject to biasing or corrupting commercial pressures [97]. A healthy democratic public sphere might, in a broader sense, be an effective insulator against disinformation. But there is no obvious mechanism for how narrow and limited interventions could bring about this broad and general benefit; not least because public funding for media is an increasingly contested issue with no prospect of short-term consensus in the face of problems such as disinformation [98].

Given the lack of evidence and plausibility, our effectiveness ratings for fighting fire with water interventions are very low:

- Narrow effectiveness: 1.
- Broad effectiveness: 0.
- Total score: 0.5.

4.4.3 Nudging

Nudging, a technique of soft behavioral manipulation through exploitation of cognitive biases, is known well beyond academia thanks to successful popularizations such as Thaler and Sunstein's "Nudge" [99]. Nudging in and of itself is probably somewhat overhyped, given that it generally only works in simplistic choice situation [100] and that the observed experimental effectiveness of different nudges varies widely [101]. Nonetheless, there is evidence that certain types of nudges could be effective for reducing user interactions with disinformation.

A team of researchers has found a consistent effect of accuracy nudges across multiple studies [102, 103, 104]. Accuracy nudges are prompts on social media posts that ask users to think about whether the information they see is accurate. Importantly, however, an independent replication of the original accuracy nudge study by said research team [105] failed to produce the same positive effects [106], indicating that the research team in question

might be introducing some amount of unintended motivated reasoning and confirmation bias into their results. Accuracy nudges, if we provisionally accept that they are potentially effective, do not improve cognition in the sense of more deliberate thinking within the dual-process account of reasoning [107]. Instead, they seem to change what people think about while deliberating [108].

Another type of nudge that has proven effective in an experimental setting is information about social norms. More specifically, injunctive social norm messages that describe what behavior most people approve of increase users' willingness to report misinformation [109].

It is unclear how well such nudges work in a broader context. In order for them to have significant real-world impact, social media platforms would have to use them on a large scale, perhaps even as a default prompt on all content. It seems unrealistic that any platform would do so, and it is unknown whether the nudges would still be effective once they were omnipresent and users got used to them. For that reason, we rate the broad effectiveness of nudging interventions lower than their narrow effectiveness:

- Narrow effectiveness: 3.
- Broad effectiveness: 0.
- Total score: 1.5.

4.4.4 Fact-checking

In their review of disinformation interventions, Courchesne et al. [14] find that fact-checking interventions are the most studied intervention type in the context of disinformation. Three reviews of research on fact-checking conclude that fact-checking can change beliefs, but that the effect is generally weak, with many studies finding no effect, and there is strong persistence of misinformation-related beliefs in spite of corrections in the form of fact-checking [110, 111, 112]. A recent meta-analysis that focused specifically on health misinformation disseminated on social media found that fact-checking has a significant positive effect on beliefs [113]. A recent study conducted in four countries concludes that fact-checking interventions successfully decrease belief in misinformation [114].

The persistence of misinformation-related beliefs has been recognized as a major limitation of and challenge for fact-checking interventions. Both cognitive (such as a failure to integrate corrections in relevant mental models), as well as socio-affective factors (such as perceiving corrections as threats to one's worldview), can be barriers to fact-checking based belief revision [115].

These limitations demonstrate that fact-checking is a complex undertaking. Correcting misinformation and providing more accurate information is an essential part of rational discourse that is, on some level, obviously effective – human civilization as a whole has managed to update and revise a myriad of incorrect beliefs over the course of our history. Yet, in a more localized context, engaging in corrections does not automatically lead to more informed and rational beliefs, and there is no guarantee that collective beliefs only shift towards a closer match with reality; epistemic backslides are possible. And not only is the process of real-world belief formation complex. The process of real-world fact-checking or debunking itself is not always as straightforward as it is in simplified experimental settings [116]. Misinformation and disinformation are often not simple black-or-white claims but rather a set of propositions that contain some elements grounded in reality. This makes accurate and precise corrections difficult. Additionally, the sheer volume of misinformation and disinformation is almost impossible to adequately and comprehensively correct. Fact-checking is always a limited, selective response to only a fraction of all the misinformation and disinformation that is circulating. These conceptual caveats in combination with the mixed empirical evidence leads us to rating the effectiveness of fact-checking interventions cautiously:

- Narrow effectiveness: 2.
- Broad effectiveness: 1.
- Total score: 1.5.

4.4.5 Prebunking

Prebunking is essentially debunking or fact-checking that is applied preemptively before an individual has come into contact with misinformation or disinformation. More specifically, prebunking is a process of psychological inoculation that consists of a warning that activates threat perception in order to motivate resistance, and a refutational preemption in the form of a substantive fact-checking [117].

There are two subtypes of prebunking interventions: Fact-based and logic-based prebunking. Fact-based prebunking is focused on preemptively correcting specific factual claims, whereas logic-based prebunking is focused on preemptively educating about the general nature of misinformation and disinformation, such as common fallacious reasoning or manipulative techniques. Both fact-based [118, 119, 120, 121] and logic-based [122, 123, 124, 125, 126] prebunking has been found to be effective at inoculating against misinformation in experimental settings. Logic-based prebunking has been found

to have the added benefit of conferring a generalized blanket of protection [121, 127]: Logic-based prebunking increases resilience against misinformation and disinformation in general, regardless of the specific topics at hand.

Only few studies have directly compared the effects of debunking and prebunking interventions. One experimental study found that debunking and fact-based prebunking have similar efficacy [128], whereas another experimental study found that fact-based prebunking has lower efficacy than debunking [129].

Overall, the literature suggests that prebunking is a highly effective intervention that potentially avoids the problem of persistence of false beliefs that is present in the context of fact-checking. Since prebunking is applied before incorrect beliefs are formed, it avoids and prevents the problem. Promising though prebunking sounds, it has at least two major practical limitations. First, prebunking interventions are much more resource intensive than debunking interventions. Prebunking interventions need to be planned, prepared, and successfully deployed before a problem arises. Doing this requires significant foresight and labor. Second, it is unclear how prebunking interventions can be scaled for real-world impact. For example, fact-checking has become a widespread practice in journalism that reaches large audiences [130], not least because it is very much compatible with journalistic practice. Fact-checking can be a part of captivating journalistic narratives and storytelling, whereas prebunking as a much more educational measure, especially in the case of logic-based prebunking, seems like a rather bland affair. Prebunking could be introduced through other vectors, such as government-sponsored programs in public schools, but such options are currently entirely speculative.

The high narrow impact and uncertain broad impact of prebunking interventions is reflected in our effectiveness ratings:

- Narrow effectiveness: 3.
- Broad effectiveness: 0.
- Total score: 1.5.

4.5 Summary

In this section, we have discussed the evidence and plausibility of the effectiveness of the thirteen disinformation interventions described in our taxonomy in section 3. Our numeric effectiveness ratings are summarized in Table 5.

Table 5: *Effectiveness of disinformation interventions.*

Intervention locus	Intervention type	Effectiveness
Sender	Blocking	4
	Deplatforming	3
	Forcing verification	3.5
	Reducing source visibility	2.5
	Labeling sources	2
Content	Deleting content	1.5
	Reducing content visibility	1.5
	Labeling content	0.5
Recipient	Fire with fire	4
	Fire with water	0.5
	Nudging	1.5
	Fact-checking	1.5
	Prebunking	1.5

5 Intervention harm

In this section, we evaluate the potential harm that disinformation intervention types cause. The harm risk estimates are not based on existing literature (there is little research on this question) but instead on a particular normative view of democratic theory: Deliberative democracy.

5.1 Defining harm

In the introduction, we have briefly outlined the ethical rationale for taking potential harm of disinformation interventions seriously: Interventions are harmful when they reduce deliberative pluralism and freedom of speech. That view is rooted in the perspective of deliberative democracy [131, 132]. Deliberative views of democracy stress the importance of rational deliberation taking place in the public sphere. Disinformation works by exploiting the openness and pluralism that make such deliberation possible. Interventions against disinformation, we argue, are harmful when they, intentionally or unintentionally, reduce the very openness and pluralism of democratic deliberation.

More specifically, we define intervention harm as damage to the principles of the Habermasian ideal speech situation. Jürgen Habermas describes an ideal speech situation as a discursive constellation with four properties [133, 45-49]:

- *Openness and inclusion*: No potential participants with opinions relevant to the discussion should be excluded.
- *Communicative equality*: All participants have equal opportunity to state their opinions.
- *No deceptions*: Participants should sincerely believe what they are saying.
- *No coercion*: There are no impediments to the free flow of arguments.

Of course, such an ideal speech situation has probably never existed in reality. But that is not the point here. Disinformation interventions can potentially damage actually existing speech situations, and the Habermasian concept of the ideal speech situation is merely an analytical tool for understanding how. In other words: We conceptualize harm as a relative decrease in deliberative quality, and the dimensions of the ideal speech situation describe the dimension on which that decrease can take place. It is important to note here that deliberative harm does not occur if disinformation is removed with surgical precision. Disinformation itself is a form of deliberative harm (it is deceptive), and removing it is beneficial. The problem is that disinformation interventions mostly lack surgical precision and cause collateral damage.

One important conceptual aspect of our harm estimates is the distinction between disinformation and misinformation. Disinformation, as argued before, is a subtype of misinformation: When a malicious actor knowingly deploys misinformation in order to achieve some goal, we are faced with disinformation. In practice, however, this question of intent is often difficult or even impossible to determine. Misinformation can be identified by its very propositional content. But it is very difficult to assess whether an actor who is disseminating misinformation is doing so knowingly or not. Typically, this assessment can only be made with high confidence by uncovering clandestine influence operations that cannot plausibly be assumed to be expressions of genuine and authentic beliefs.

In the context of deliberative principles, the distinction between misinformation and disinformation and the question of intent are crucial. In democratic debate, genuine misinformation generally has to be tolerated, since speech participants have the right to be irrational; after all, the very ambition of deliberation is to procedurally weed out bad arguments and arrive at a consensus with good arguments. But deceptive, fake disinformation, as noted in the concept of the ideal speech situation, does not have to be tolerated, since it is an attack on genuine deliberation. The practical problem in the context of disinformation interventions is that disinformation cannot

be reliably detected. From an ethical perspective, this leads to a cautionary principle: Without evidence that strongly suggests otherwise, we have to assume that misinformation is indeed genuine, authentic misinformation and not inauthentic, manipulative disinformation. We have to, in other words, apply a sort of principle of charity: When in doubt, we have to assume that we are dealing with misinformation rather than disinformation.

This cautionary principle means a bias towards false negatives and against false positives. We regard interventions that successfully reduce such misinformation for which there is no strong indication of malicious and conscious intent as potentially harmful because they curtail deliberative pluralism and freedom of speech. We discuss this harm conundrum below for each intervention type where it occurs.

Our ratings of intervention harm are similar to those for effectiveness. We rate intervention harm on a scale of 0 (no harm) to 4 (high harm). For the effectiveness ratings, we rated two subscales with increments of 1, and the final score has increments of 0.5. For the harm risk ratings, there is only one score which we rate in increments of 0.5 in order to have a scale comparable to that of the effectiveness ratings.

5.2 Sender interventions

5.2.1 Blocking

Blocking means categorically excluding some sender from a public debate. Categorical exclusion is a highly effective intervention, but it is also a very risky one. In order for categorical exclusions to be morally just, it has to be demonstrated that an actor is only or mainly knowingly communicating in bad faith in order to achieve some goal by doing so. That is a very high evidentiary bar to clear, so we rate the harm risk of blocking as very high: 4.

5.2.2 Deplatforming

The harm risk of deplatforming is similar to the risk of blocking, only on a smaller scale. In order to justifiably prevent an actor from communicating in a certain communication vector, there needs to be strong evidence that the actor in question is only or mainly intentionally disseminating disinformation. Given the difficulty of proving such intent, the risk of false positives – deplatforming spreaders of misinformation who are genuinely irrational and not deceptive – is high. Accordingly, our harm risk rating is high: 3.

5.2.3 Forcing verification

Mandatory user identity verification would significantly reduce malicious actors' ability to use fake user profiles and social bots for spreading disinformation. However, mandatory verification also carries some deliberative risks by reducing inclusion and equality and by introducing a form of communicative coercion [134]. Anonymity can be beneficial in debates about stigmatized or taboo topics, and it can be a layer of protection against persecution and sanctions for vulnerable individuals and groups, such as whistleblowers or political dissidents. Forced identity verification could in some cases hamper the flow of free and critical arguments and instead induce communication that is more deferential to power holders. Overall, we estimate that the harm risk of forcing verification is high: 3.

5.2.4 Reducing source visibility

The harm risk of reducing source visibility is similar to that of deplatforming. In both cases, the evidentiary requirement for a justifiable intervention is high, as is the probability of false positives. However, given that reducing source visibility is a less severe intervention than deplatforming, our harm risk rating is lower: 2.5.

5.2.5 Labeling sources

Adding labels to accounts and users on social media is in principle merely a transparency measure that does not present any immediate deliberative risks. In a broader sense, however, the question is which sources receive what kind of label. So far, social media platforms have applied labels only selectively and sparingly. This creates the risk of biased labeling. If, for example, a platform applies the "state-affiliated media" label only selectively, that could lend undue credibility to other similar media outlets who happen not to receive such a label. This could indirectly reduce communicative equality. Overall, however, we estimate the harm risk of labeling sources to be rather low: 1.5.

5.3 Content interventions

5.3.1 Deleting content

The main harm risk of deleting content are false positives. While misinformation can be spotted somewhat accurately, reliably inferring malicious intent and therefore classifying misinformation as intentional disinformation is far

more difficult. The risk here is that, in order to delete a meaningful amount of disinformation, a large volume of merely genuine, authentically believed misinformation would have to be deleted. From a deliberative point of view, that reduction of communicative equality is a high price to pay. Therefore, our harm risk estimate for deleting content is high: 3.

5.3.2 Reducing content visibility

The harm risk of reducing content visibility is similar to the harm risk of deleting content. Reducing content visibility negatively impacts communicative equality, given that a meaningful effect against disinfromation would require a significant proportion of false positives. Given that the content is merely less visible and not outright deleted, our harm risk score is lower than for deleting content: 2.5.

5.3.3 Labeling content

Neither veracity labeling nor contextual labeling, the two main subtypes of the labeling content intervention, are inherently problematic from a deliberative point of view. Such interventions could even be seen to improve the quality of the communication by pointing out inaccuracies or providing further relevant arguments. Indirectly, the question arises how accurate and precise such labels themselves are, and how universally they are deployed. Similar to the labeling sources intervention, a lack of labeling on content that is in fact dubious could lend that content undue credibility. Overall, we estimate the harm risk of labeling content to be low, especially in the case of contextual labels: 1.

5.4 Recipient interventions

5.4.1 Fire with fire

Using disinformation in order to fight disinformation is quite obviously problematic from a deliberative point of view. Disinformation is deceptive in nature and a sabotage of genuine discourse. We therefore estimate the harm risk of the fighting fire with fire intervention to be very high: 4.

5.4.2 Fire with water

The fighting fire with water intervention poses little harm risk. Adding relevant sources to the deliberative process is generally beneficial to the quality of the discourse. One slight source of risk is the question of who creates the water

that is to fight the fire. The enrichment of the information environment with high-quality sources and content should ideally be an organic process such as a well-functioning system of independent journalistic media. Overall, we estimate the harm risk of fighting fire with water to be very low: 0.5.

5.4.3 Nudging

Proponents of nudging argue that nudging is generally not manipulative because nudging neither removes choice options nor forces people into certain choices [135, 136]. But it is difficult to regard nudging as non-manipulative seeing that the whole point of nudging is to make people behave in a desired way [137]. Even if the goal behavior is beneficial to the people affected by nudging, the intervention remains manipulative.

The manipulative nature of nudging interventions is potentially harmful from a deliberative perspective in two ways. First, nudges are inherently deceptive. They are interventions that work because people who are affected by them do not realize what they are being exposed to. Second, given the deceptive nature of nudging, nudges risk creating limited local optima at the cost of reducing the probability of global optima [138]. In a given decision-making situation, a nudge can make a good outcome in that decision-making situation more probable (local optimum). But since nudges bypass rational cognition, people exposed to nudges do not learn from that experience and they do not become more resilient against disinformation in general (global optimum). This means that, all else being equal, a disinformation intervention that stimulates active, rational cognition and potentially increases individual resilience against disinformation is always preferable to nudging, since nudging can only achieve local optima, whereas non-deceptive, cognition-based interventions can achive both local and global optima.

Given the manipulative nature of nudges, we rate the harm risk of this intervention to be moderate: 2.

5.4.4 Fact-checking

In general, fact-checking does not cause deliberative harm because it is merely a contribution of arguments to a debate. Of course, it is possible, as has been criticized before [139], that fact-checking efforts themselves are flawed[3]. But if that is the case, pointing out and criticizing such flaws is simply part of

[3]This argument should not suggest that journalistic fact-checking is generally flawed. The evidence points to the opposite: Fact-checking is generally done transparently and carefully [140].

the deliberative process. Overall, we estimate that fact-checking has no harm risk and rate it accordingly: 0.

5.4.5 Prebunking

The general harm risk of prebunking is comparable to that of fact-checking. Prebunking does not violate any of the principles of deliberative speech. One potential indirect risk of popularizing the use of prebunking is that the technique can be used to inoculate against correct arguments in order to promote disinformation. Psychological inoculation is cause-neutral: It is a technique for conferring resistance to persuasion attempts [141]. That resistance can in theory also be resistance against good arguments.

But this slight risk of using prebunking to inoculate against correct information is not a direct consequence of using prebunking as a disinformation intervention. We therefore estimate that prebunking has no harm risk and give it a corresponding rating: 0.

5.5 Summary

In this section, we have discussed the potential harm of the thirteen disinformation interventions from a deliberative perspective. Our numeric harm risk ratings are summarized in Table 6.

Table 6: *Harm risks of disinformation interventions.*

Intervention locus	Intervention type	Harm risk
Sender	Blocking	4
	Deplatforming	3
	Forcing verification	3
	Reducing source visibility	2.5
	Labeling sources	1.5
Content	Deleting content	3
	Reducing content visibility	2.5
	Labeling content	1
Recipient	Fire with fire	4
	Fire with water	0.5
	Nudging	2
	Fact-checking	0
	Prebunking	0

6 Intervention ethics

In this section, we combine our analyses on intervention effectiveness and intervention harm risk into our proposed ethical framework. The framework consists of two analytical steps. We first perform an analysis of overall intervention benefits in order to reveal unobjectionable interventions that have positive net benefits. Second, we identify interventions with high effectiveness and high harm potential which, despite their prima facie acceptable net benefit of around zero, are ethically problematic given the potentially large amount of damage they can cause.

6.1 Symmetrical evaluation: Net benefits

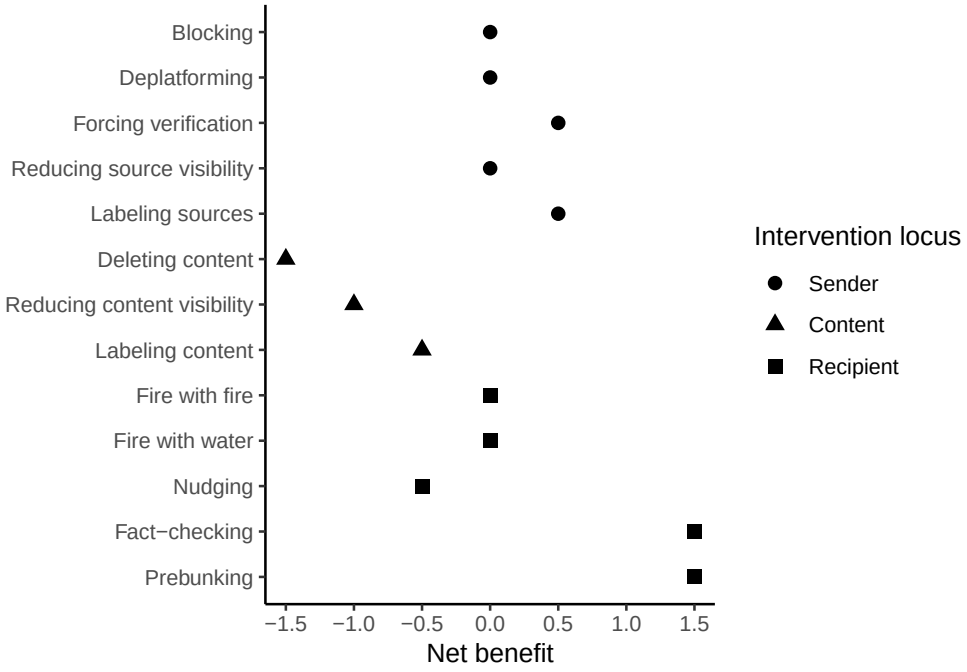
The first step of our ethical analysis is the simple depiction of total net benefits of interventions. In this analytical step, which assumes symmetry between benefit and harm, the focus is on interventions that pass the threshold towards a clear net benefit. As an ethical heuristic, a positive net benefit is generally indicative of interventions that are safe to deploy. The symmetrical ethical evaluation is depicted in Figure 6. The net benefits are calculated by subtracting the harm scores from the effectiveness scores.

Some interventions, as is visible in the chart, create overall disvalue: The potential harms of deleting content, reducing content visibility, labeling content, and nudging are negative. A number of interventions have a net zero benefit: Blocking, deplatforming, reducing source visibility, fighting fire with fire, and fighting fire with water all do as much potential harm as they do potential good. Four interventions have a positive net impact: Forcing verification, labeling sources, fact-checking, and prebunking. Fact-checking and prebunking have the most favorable net benefit, whereas deleting content has the most unfavorable net benefit. The group of content-focused interventions fares worst.

6.2 Asymmetrical evaluation: Acceptable harm

The second step of the ethical analysis shifts the attention to the relationship between effectiveness and harm. The four quadrants we propose can be used as a guide for evaluating interventions with greater precision. Within this analytical tool, interventions with high effectiveness and low harm are most favorable, and those with low effectiveness and high harm are least favorable. The most demanding intervention quadrant is comprised of interventions that have high effectiveness but also do a lot of harm. For interventions in this quadrant, we should assume an asymmetry between the weight of

Figure 6: *Net benefits of disinformation interventions.*



effectiveness and the weight of potential harm: The harm they potentially do is worse than the same amount of good they might do. For that reason, use of those interventions is justifiable when and only when they can be used with increased precision so that collateral damage in the form of false positives can be avoided.

The asymmetrical evaluation of disinformation effectiveness is depicted in Figure 7. The scatterplot shows effectiveness plotted against harm.

The scatterplot consists of four quadrants. In the lower left is the low effectiveness and low harm quadrant. These are interventions that are generally democratically safe to use, but they are not necessarily useful. The ethically most favorable interventions in this quadrant are fact-checking and prebunking. Those interventions have limited effectiveness, but they do no harm.

In the lower right is the quadrant of high effectiveness and low harm interventions. From an ethical point of view, this is the most desirable quadrant: The only intervention that is narrowly in this quadrant, labeling sources, is fairly effective while doing comparatively little harm.

In the upper left is the low effectiveness and high harm quadrant. Interven-

tions in this quadrant are generally unfavorable because they do potentially significant harm while having limited effectiveness. This ethical assessment holds even if we posit symmetrical weight of harm and effectiveness.

In the upper right is the high effectiveness and high harm quadrant. Interventions in this quadrant are the most problematic. They have both high positive effectiveness and high negative impact. This is where an asymmetrical view of effectiveness and harm comes into play. Even though the net benefit of those interventions is around zero, the large amount of deliberative harm they potentially cause means that their use should be restricted. In order for such interventions to be acceptable, there needs to be a strong justification for the particular use case in question. That justification needs to include a realistic plan for increasing the precision of the interventions so that the potential harm is reduced and fewer false positives result.

7 Discussion

Given the alarming developments of recent years, disinformation is likely to become an ever more pressing problem in the future. Various actors – governments, tech platforms, civil society organizations, researchers, and others – are scrambling to find effective ways of addressing the threat. In this race against disinformation, however, it is important to keep the other side of the equation in mind: Disinformation interventions can do good by pushing back against disinformation, but they can also do harm by damaging the very principles and rules of democratic deliberation that disinformation itself seeks to sabotage. It is crucial that democratic societies defend themselves against disinformation – but, as we argue in this paper, not at any price.

7.1 Unknown unknowns and future directions

There is a lot of uncertainty about disinformation. Despite large efforts to detect and curb disinformation, we do not know who is sending what kind of and how much disinformation through which communication vectors to which audiences with what kinds of effects. This uncertainty exists because disinformation is to a large degree a clandestine, deceptive effort. The unfortunate consequence of disinformation's clandestine nature is that the overall problem has the properties of an unknown unknown [142]. When it comes to disinformation, we do not know the true extent of the problem, and we do not know what exactly it is we do not know.

This leads to a second problem, but at least this one has the quality of a known unknown. Given the slippery nature of disinformation, the evidence

on the effectiveness of disinformation interventions is limited. This is the main limitation of our study: Our effectiveness ratings might convey a sense of certainty, given that they formally look like precise point estimates. But it is epistemologically more appropriate to think of our effectiveness ratings as estimates with fairly broad posterior distributions. We do believe that the available evidence and logical plausibility point towards certain levels of effectiveness, but we are still in the early stages of understanding the real-world impact of disinformation interventions [13]. Empirical research should address the question of broad, real-world impact more. Laboratory experimental studies are valuable, but in a complex problem like disinformation, narrow experimental efficacy does not necessarily translate into broad real-world effectiveness.

The effectiveness ratings in this paper should therefore not be regarded as a definitive answer to the question of effectiveness, but instead as a preliminary analysis that can and should be criticized and updated. The ethical framework we propose, consisting of a combined analysis of effectiveness and deliberative harm, can serve as an analytical foundation for this kind of future inquiry.

Another aspect that should be tackled in future research is the question of intervention effectiveness and harm in the context of misinformation as well as mal-information such as hate speech [7]. Even though the intervention types are either the same (in the case of misinformation) or at least have significant overlap (in the case of hate speech) in these three domains, the logic of their evaluation might differ. For example, deplatforming a known disinformation agent has a different ethical implication than deplatforming a person who merely shares genuine misinformation. Our very argument that disinformation interventions can be harmful rests precisely on the premise that silencing disinformation agents is ethically desirable, but false-positives in the form of silencing genuine misinformation believers is not. Similarly, the ethical principles of hate speech interventions need to be analyzed separately as well. Deplatforming a person who is spreading hate speech, for example, might represent a curtailing of the deliberative right to spread genuine mal-information, but from an ethical perspective, hate speech, genuine though it may be, is arguably less acceptable than misinformation because the latter does not directly represent attacks against vulnerable individuals and groups.

7.2 The question of intervention source

Let us compare two fictional scenarios. In scenario A, managers at a social media company decide to deplatform a certain account for posting disinformation. In scenario B, the government forces the social media company to deplatform that account for the exact same reason and with the exact same

evidentiary basis. Is there, ethically, a difference between scenarios A and B?

A typical deontological intuition here would be that yes, there is indeed a difference. The common view of such scenarios is that government action is more problematic than private action because speech is typically protected against government action whereas there are no legal rights to being able to speak in a private forum [143]. From a consequentialist view, however, there is no moral difference between the scenarios. In both cases, the moral status of the intervention hinges on whether the deplatforming decision is really justified. If the banned account is indeed a malicious actor who only or mostly engages in knowingly disseminating disinformation, the intervention is beneficial. If the banned account is not in fact a malicious actor, the intervention has caused deliberative harm.

The argument of this paper is that, in order to ethically evaluate disinformation interventions, a focus on consequences rather than on intervention source is more useful. If, for example, a person is unjustly banned from major social media platforms for allegedly spreading disinformation (a false positive), the person in question cannot access important fora of public discourse any longer. That exclusion obviously constitutes deliberative damage. Whether the wrong deplatforming decisions were made by private entities or by the government is, literally, of no consequence – the amount of damage is the same. Conversely, applying deontological standards to government efforts against disinformation is counterproductive. Most interventions have non-zero risk of causing deliberative harm, which makes them categorically unacceptable from a deontological perspective. That is an irrational ethical stance. Categorically rejecting interventions with favorable net benefits because they can cause small amounts of harm is ultimately little more than a form of omission bias [144].

Overall, the moral status of disinformation interventions is not contingent on who initiates them but on what effects and consequences they have. Only such a consequentialist allows us to perform meaningful ethical analysis.

7.3 Policy recommendations

What does our analysis mean for public and private actors who engage in interventions against disinformation? We propose three rules of principles that should guide the use of disinformation interventions.

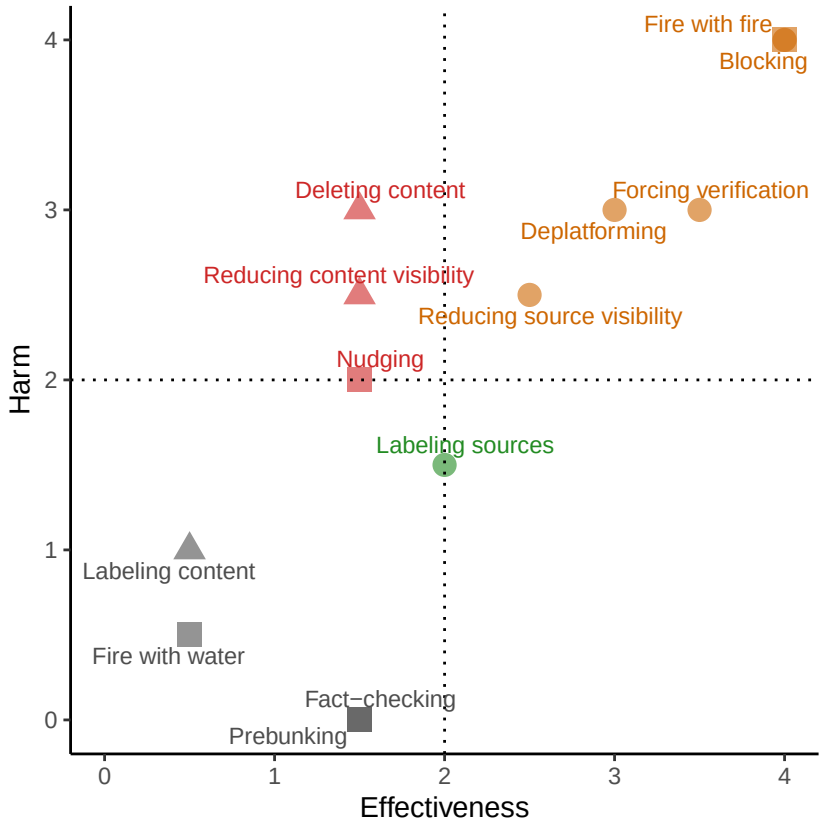
First, the *cautionary principle* should be applied. Not all misinformation is disinformation, and misinformation is an aspect of democratic deliberation that has to be tolerated – participants in democratic discourse have the right to hold irrational beliefs. The cautionary principle means that false positives have to be avoided: Disinformation interventions should not target genuine

misinformation and the people who believe in it. Given how difficult it is to determine intent, the default assumption in unclear cases is that an actor is not acting (only or mainly) maliciously and deceptively but that they are genuinely believing what they are saying.

Second, priority should be given to interventions with favorable overall *net benefit*. In our analysis depicted in Figure 6, the three interventions with highest net benefit are fact-checking (debunking), prebunking, and labeling sources. This evaluation can of course change with new data. What matters is the underlying principle of net benefit calculated by subtracting harm from effectiveness.

Third, the use of *high impact and high harm interventions* as depicted in Figure 7 requires special justification. Those are interventions that have an overall net benefit of around zero. In those cases, however, an asymmetric view of harm and effectiveness is more appropriate: The high amount of deliberative harm those interventions do generally makes them unacceptable. High impact and high harm interventions can be thought of, to use a crude kinetic warfare analogy, as weapons of mass destruction which are generally unacceptable given the massive collateral damage they cause. If such disinformation interventions are to be deployed, their use must be justified by substantive arguments and evidence that they can be used in a precise manner that significantly reduces the amount of deliberative harm they cause.

Figure 7: *Intervention effectiveness plotted against intervention harm.*



Note: Circles are sender-focused interventions; triangles are content-focused interventions; squares are recipient-focused interventions. The colors represent the position in the four quadrants. The data points for prebunking and fact-checking (lower left quadrant) and blocking and fire with fire (upper right quadrant) overlap.

References

[1] Christopher Paul and Miriam Matthews. The Russian "Firehose of Falsehood" Propaganda Model. Technical report, RAND Corporation, July 2016.

[2] Ilya Yablokov. Conspiracy Theories as a Russian Public Diplomacy Tool: The Case of Russia Today (RT). *Politics*, 35(3-4):301–315, November 2015.

[3] Charlotte Wagnsson. The paperboys of Russian messaging: RT/Sputnik audiences as vehicles for malign information influence. *Information, Communication & Society*, 0(0):1–19, February 2022.

[4] Philipp Müller and Anne Schulz. Alternative media for a populist audience? Exploring political and media use predictors of exposure to Breitbart, Sputnik, and Co. *Information, Communication & Society*, 24(2):277–293, January 2021.

[5] James Pamment. How the Kremlin circumvented EU sanctions on Russian state media in the first weeks of the illegal invasion of Ukraine. *Place Branding and Public Diplomacy*, September 2022.

[6] Aiden Hoyle and Peter B. M. J. Pijpers. Stemming the Narrative Flow: The Legal and Psychological Grounding for the European Union's Ban on Russian State-Sponsored Media, September 2022.

[7] Claire Wardle and Hossein Derakhshan. Information disorder: Toward an interdisciplinary framework for research and policymaking. Council of Europe report DGI(2017)09, Council of Europe, Strasbourg, 2017.

[8] Deen Freelon and Chris Wells. Disinformation as Political Communication. *Political Communication*, 37(2):145–156, March 2020. Publisher: Routledge _eprint: https://doi.org/10.1080/10584609.2020.1723755.

[9] Chris Tenove. Protecting Democracy from Disinformation: Normative Threats and Policy Responses. *The International Journal of Press/Politics*, 25(3):517–537, July 2020.

[10] Spencer McKay and Chris Tenove. Disinformation as a Threat to Deliberative Democracy. *Political Research Quarterly*, 74(3):703–717, September 2021.

[11] Nadya Bliss, Elizabeth Bradley, Joshua Garland, Filippo Menczer, Scott W. Ruston, Kate Starbird, and Chris Wiggins. An Agenda for Disinformation Research, December 2020. arXiv:2012.08572 [cs].

[12] Anastasia Kozyreva, Philipp Lorenz-Spreen, Stefan Herzog, Ullrich Ecker, Stephan Lewandowsky, and Ralph Hertwig. Toolbox of Interventions Against Online Misinformation and Manipulation, December 2022.

[13] Sacha Altay. How Effective Are Interventions Against Misinformation?, May 2022.

[14] Laura Courchesne, Julia Ilhardt, and Jacob N. Shapiro. Review of social science research on the impact of countermeasures against influence operations. *Harvard Kennedy School Misinformation Review*, September 2021.

[15] Jesús Aguerri, Mario Santisteban, and Fernando Miró-Llinares. The fight against disinformation and its consequences: Measuring the impact of "Russia state-affiliated media" on Twitter, April 2022.

[16] Anastasia Kozyreva, Stefan Herzog, Stephan Lewandowsky, Ralph Hertwig, Philipp Lorenz-Spreen, Mark Leiser, and Jason Reifler. Free speech vs. harmful misinformation: Moral dilemmas in online content moderation, June 2022.

[17] Corneliu Bjola. The Ethics of Countering Digital Propaganda. *Ethics & International Affairs*, 32(3):305–315, 2018.

[18] Larry Alexander and Michael Moore. Deontological Ethics. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2021 edition, 2021.

[19] Walter Sinnott-Armstrong. Consequentialism. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2022 edition, 2022.

[20] Magnus Vinding. *Suffering-Focused Ethics: Defense and Implications*. Ratio Ethica, May 2020.

[21] Julie Posetti and Alice Matthews. A short guide to the history of 'fake news' and disinformation. Technical report, International Center for Journalists, Washington, D.C., 2018.

[22] Samantha Bradshaw and Philip N. Howard. The Global Organization of Social Media Disinformation Campaigns. *Journal of International Affairs*, 71(1.5):23–32, 2018.

[23] Samantha Bradshaw, Hannah Bailey, and Philip N. Howard. Industrialized Disinformation: 2020 Global Inventory of Organized Social Media Manipulation. Working Paper 2021.1, Computational Propaganda Project at the Oxford Internet Institute, Oxford, UK, 2021.

[24] Philipp Lorenz-Spreen, Lisa Oswald, Stephan Lewandowsky, and Ralph Hertwig. A systematic review of worldwide causal and correlational evidence on digital media and democracy. *Nature Human Behaviour*, pages 1–28, November 2022.

[25] Countering Disinformation through Democracy and Governance Programming: Protecting the Integrity of Political and Electoral Information and Discourse, 2021.

[26] Kamya Yadav, Ulaş Erdoğdu, Samikshya Siwakoti, Jacob N. Shapiro, and Alicia Wanless. Countries have more than 100 laws on the books to combat misinformation. How well do they work? *Bulletin of the Atomic Scientists*, 77(3):124–128, May 2021.

[27] Corneliu Bjola and James Pamment. Digital containment: Revisiting containment strategy in the digital age. *Global Affairs*, 2(2):131–142, March 2016.

[28] Alberto Alemanno. How to Counter Fake News? A Taxonomy of Anti-fake News Approaches. *European Journal of Risk Regulation*, 9(1):1–5, March 2018.

[29] Kamya Yadav. Platform Interventions: How Social Media Counters Influence Operations. PCIO Baseline, Carnegie Endowment for International Peace, Washington, D.C., January 2021.

[30] Kai Shu, Amrita Bhattacharjee, Faisal Alatawi, Tahora H. Nazer, Kaize Ding, Mansooreh Karami, and Huan Liu. Combating disinformation in a social media age. *WIREs Data Mining and Knowledge Discovery*, 10(6):e1385, 2020.

[31] Harold D. Lasswell. The structure and function of communication in society. *The communication of ideas*, 37(1):136–139, 1948.

[32] Natascha A. Karlova and Karen E. Fisher. A social diffusion model of misinformation and disinformation for understanding human information behaviour. *Information Research*, 18(1), March 2013.

[33] Stacey Cone. Presuming A Right to Deceive. *Journalism History*, 24(4):148–156, January 1999.

[34] Cass R. Sunstein. Nudging: A Very Short Guide. *Journal of Consumer Policy*, 37(4):583–588, December 2014.

[35] John A. Banas and Stephen A. Rains. A Meta-Analysis of Research on Inoculation Theory. *Communication Monographs*, 77(3):281–311, September 2010.

[36] Michelle A. Amazeen, Fabrício Benevenuto, Nadia M. Brashier, Robert M. Bond, Lia C. Bozarth, Ceren Budak, Ullrich K. Ecker, Lisa K. Fazio, Emilio Ferrara, Andrew J. Flanagin, Alessandro Flammini, Deen Freelon, Nir Grinberg, Ralph Hertwig, Kathleen Hall Jamieson, Kenneth Joseph, Jason J. Jones, R. Kelly Garrett, Daniel Kreiss, Shannon McGregor, Jasmine McNealy, Drew Margolin, Alice Marwick, FiIippo Menczer, Miriam J. Metzger, Seungahn Nah, Stephan Lewandowsky, Philipp Lorenz-Spreen, Pablo Ortellado, Irene Pasquetto, Gordon Pennycook, Ethan Porter, David G. Rand, Ronald Robertson, Briony Swire-Thompson, Francesca Tripodi, Soroush Vosoughi, Chris Vargo, Onur Varol, Brian E. Weeks, John Wihbey, Thomas J. Wood, and Kai-Cheng & Yang. Tackling misinformation: What researchers could do with social media data. *Harvard Kennedy School Misinformation Review*, 1(8), November 2020.

[37] Sonali Chandel, Zang Jingji, Yu Yunnan, Sun Jingyao, and Zhang Zhipeng. The Golden Shield Project of China: A Decade Later—An in-Depth Study of the Great Firewall. In *2019 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)*, pages 111–119, October 2019.

[38] Chong Zhang. Who bypasses the Great Firewall in China? *First Monday*, March 2020.

[39] Margaret E. Roberts. *Censored: distraction and diversion inside China's great firewall*. Princeton University Press, Princeton, New Jersey, 2018.

[40] Sara Bundtzen and Mauritius Dorn. Banning RT and Sputnik Across Europe: What Does it Hold for the Future of Platform Regulation? Technical report, Institute for Strategic Dialogue, London, April 2022.

[41] Elise Thomas. Why Western conspiracy influencers are promoting pro-Kremlin propaganda. Technical report, Institute for Strategic Dialogue, London, March 2022.

[42] Marko Kovic, Adrian Rauchfleisch, Marc Sele, and Christian Caspar. Digital astroturfing in politics: Definition, typology, and countermeasures. *Studies in Communication Sciences*, 18(1):69–85, November 2018. Number: 1.

[43] N. F. Johnson, R. Leahy, N. Johnson Restrepo, N. Velasquez, M. Zheng, P. Manrique, P. Devkota, and S. Wuchty. Hidden resilience and adaptive dynamics of the global online hate ecology. *Nature*, 573(7773):261–265, September 2019.

[44] Shagun Jhaver, Christian Boylston, Diyi Yang, and Amy Bruckman. Evaluating the Effectiveness of Deplatforming as a Moderation Strategy on Twitter. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):381:1–381:30, October 2021.

[45] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):31:1–31:22, December 2017.

[46] Joan Donovan, Becca Lewis, and Brian Friedberg. Parallel Ports. Sociotechnical Change from the Alt-Right to Alt-Tech. In *Parallel Ports. Sociotechnical Change from the Alt-Right to Alt-Tech*, pages 49–66. transcript Verlag, December 2018.

[47] Richard Rogers. Deplatforming: Following extreme Internet celebrities to Telegram and alternative social media. *European Journal of Communication*, 35(3):213–229, June 2020.

[48] Aleksandra Urman and Stefan Katz. What they do in the shadows: examining the far-right networks on Telegram. *Information, Communication & Society*, 25(7):904–923, May 2022.

[49] Ehsan Dehghan and Ashwin Nagappa. Politicization and Radicalization of Discourses in the Alt-Tech Ecosystem: A Case Study on Gab Social. *Social Media + Society*, 8(3):20563051221113075, July 2022.

[50] Binny Mathew, Anurag Illendula, Punyajoy Saha, Soumya Sarkar, Pawan Goyal, and Animesh Mukherjee. Hate begets Hate: A Temporal

Study of Hate Speech. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):92:1–92:24, October 2020.

[51] H. Innes and M. Innes. De-platforming disinformation: conspiracy theories and their control. *Information, Communication & Society*, 0(0):1–19, October 2021.

[52] Adrian Rauchfleisch and Jonas Kaiser. Deplatforming the Far-right: An Analysis of YouTube and BitChute, June 2021.

[53] Maura Conway, Moign Khawaja, Suraj Lakhani, Jeremy Reffin, Andrew Robertson, and David Weir. Disrupting Daesh: Measuring Takedown of Online Terrorist Material and Its Impacts. *Studies in Conflict & Terrorism*, 42(1-2):141–160, February 2019.

[54] J.M. Berger and Heather Perez. The Islamic State's Diminishing Returns on Twitter: How suspensions are limiting the social networks of English-speaking ISIS supporters. Program on Extremism, George Washington University, Washington, D.C., 2016.

[55] John Suler. The Online Disinhibition Effect. *CyberPsychology & Behavior*, 7(3):321–326, June 2004.

[56] Daegon Cho and K. Hazel Kwon. The impacts of identity verification and disclosure of social cues on flaming in online user comments. *Computers in Human Behavior*, 51:363–372, October 2015.

[57] Rolf Fredheim, Alfred Moore, and John Naughton. Anonymity and Online Commenting: The Broken Windows Effect and the End of Drive-by Commenting. In *Proceedings of the ACM Web Science Conference*, WebSci '15, pages 1–8, New York, NY, USA, June 2015. Association for Computing Machinery.

[58] Maria Gruber, Christiane Mayer, and Sabine A. Einwiller. What drives people to participate in online firestorms? *Online Information Review*, 44(3):563–581, January 2020.

[59] Robert Gorwa and Douglas Guilbeault. Unpacking the Social Media Bot: A Typology to Guide Research and Policy. *Policy & Internet*, 12(2):225–248, 2020.

[60] Laura Savolainen. The shadow banning controversy: perceived governance and algorithmic folklore. *Media, Culture & Society*, 44(6):1091–1109, September 2022.

[61] Vijaya Gadde and Kayvon Beykpour. Setting the record straight on shadow banning, July 2018.

[62] Renée DiResta. The Twitter Files Are a Missed Opportunity. *The Atlantic*, December 2022.

[63] Erwan Le Merrer, Benoît Morgan, and Gilles Trédan. Setting the Record Straighter on Shadow Banning. In *IEEE INFOCOM 2021 - IEEE Conference on Computer Communications*, pages 1–10, May 2021. ISSN: 2641-9874.

[64] Kokil Jaidka, Subhayan Mukerjee, and Yphtach Lelkes. Silenced on Social Media: The Gatekeeping Functions of Shadowbans In The American Twitterverse, October 2022.

[65] Nicolas Suzor. Trying to understand YouTube's recommendation system, June 2020.

[66] Marian Friestad and Peter Wright. The Persuasion Knowledge Model: How People Cope with Persuasion Attempts. *Journal of Consumer Research*, 21(1):1–31, June 1994.

[67] Jason Ross Arnold, Alexandra Reckendorf, and Amanda L. Wintersieck. Source alerts can reduce the harms of foreign disinformation. *Harvard Kennedy School Misinformation Review*, May 2021.

[68] Jack Nassetta and Kimberly Gross. State media warning labels can counteract the effects of foreign misinformation. *Harvard Kennedy School Misinformation Review*, October 2020.

[69] Fan Liang, Qinfeng Zhu, and Gabriel Miao Li. The Effects of Flagging Propaganda Sources on News Sharing: Quasi-Experimental Evidence from Twitter. *The International Journal of Press/Politics*, page 19401612221086905, March 2022.

[70] Jadbinder Seehra, Daniel Stonehouse-Smith, and Nikolaos Pandis. Assessment of early exaggerated treatment effects in orthodontic interventions using cumulative meta-analysis. *European Journal of Orthodontics*, 43(5):601–605, October 2021.

[71] Government and state-affiliated media account labels, August 2020.

[72] Orestis Papakyriakopoulos, Juan Carlos Medina Serrano, and Simon Hegelich. The spread of COVID-19 conspiracy theories on social media and the effect of content moderation. *Harvard Kennedy School Misinformation Review*, 1(3), August 2020.

[73] Olga Papadopoulou, Evangelia Kartsounidou, and Symeon Papadopoulos. COVID-Related Misinformation Migration to BitChute and Odysee. *Future Internet*, 14(12):350, December 2022.

[74] Tarleton Gillespie. Do Not Recommend? Reduction as a Form of Content Moderation. *Social Media + Society*, 8(3):20563051221117552, July 2022.

[75] Nandita Krishnan, Jiayan Gu, Rebekah Tromble, and Lorien C. Abroms. Research note: Examining how various social media platforms have responded to COVID-19 misinformation. *Harvard Kennedy School Misinformation Review*, December 2021.

[76] Garrett Morrow, Briony Swire-Thompson, Jessica Montgomery Polny, Matthew Kopec, and John P. Wihbey. The emerging science of content labeling: Contextualizing social media content moderation. *Journal of the Association for Information Science and Technology*, 73(10):1365–1386, 2022.

[77] Alberto Ardèvol-Abreu, Patricia Delponti, and Carmen Rodríguez-Wangüemert. Intentional or inadvertent fake news sharing? Fact-checking warnings and users' interaction with social media content. *Profesional de la información*, 29(5), September 2020. Number: 5.

[78] Dongfang Gaozhao. Flagging fake news on social media: An experimental study of media consumers' identification of fake news. *Government Information Quarterly*, 38(3):101591, July 2021.

[79] Katherine Clayton, Spencer Blair, Jonathan A. Busam, Samuel Forstner, John Glance, Guy Green, Anna Kawata, Akhila Kovvuri, Jonathan Martin, Evan Morgan, Morgan Sandhu, Rachel Sang, Rachel Scholz-Bright, Austin T. Welch, Andrew G. Wolff, Amanda Zhou, and Brendan Nyhan. Real Solutions for Fake News? Measuring the Effectiveness of General Warnings and Fact-Check Tags in Reducing Belief in False Stories on Social Media. *Political Behavior*, 42(4):1073–1095, December 2020.

[80] Anne Oeldorf-Hirsch, Mike Schmierbach, Alyssa Appelman, and Michael P. Boyle. The Ineffectiveness of Fact-Checking Labels on News

Memes and Articles. *Mass Communication and Society*, 23(5):682–704, September 2020.

[81] Filipo Sharevski, Raniem Alsaadi, Peter Jachim, and Emma Pieroni. Misinformation Warning Labels: Twitter's Soft Moderation Effects on COVID-19 Vaccine Belief Echoes, April 2021. arXiv:2104.00779 [cs].

[82] Sarah E Kreps and Douglas L Kriner. The COVID-19 Infodemic and the Efficacy of Interventions Intended to Reduce Misinformation. *Public Opinion Quarterly*, 86(1):162–175, March 2022.

[83] Wallace Chipidza and Jie (Kevin) Yan. The effectiveness of flagging content belonging to prominent individuals: The case of Donald Trump on Twitter. *Journal of the Association for Information Science and Technology*, 73(11):1641–1658, 2022.

[84] Nicholas Dias, Gordon Pennycook, and David G. Rand. Emphasizing publishers does not effectively reduce susceptibility to misinformation on social media. *Harvard Kennedy School Misinformation Review*, 1(1), January 2020.

[85] Elina H. Hwang and Stephanie Lee. A Nudge to Credible Information as a Countermeasure to Misinformation: Evidence from Twitter, September 2021.

[86] Gordon Pennycook, Adam Bear, Evan T. Collins, and David G. Rand. The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Headlines Increases Perceived Accuracy of Headlines Without Warnings. *Management Science*, 66(11):4944–4957, November 2020.

[87] Holly Kathleen Hall. The new voice of America: Countering Foreign Propaganda and Disinformation Act. *First Amendment Studies*, 51(2):49–61, July 2017.

[88] A. Ross Johnson. Managing Media Influence Operations: Lessons from Radio Free Europe/Radio Liberty. *International Journal of Intelligence and CounterIntelligence*, 31(4):681–701, October 2018.

[89] Matthew Alford. The Political Impact of the Department of Defense on Hollywood Cinema. *Quarterly Review of Film and Video*, 33(4):332–347, May 2016. al.

[90] Tom Secker and Matthew Alford. New Evidence for the Surprisingly Significant Propaganda Role of the Central Intelligence Agency and

Department of Defense in the Screen Entertainment Industry. *Critical Sociology*, 45(3):347–359, May 2019.

[91] David N. Rapp and Nikita A. Salovich. Can't We Just Disregard Fake News? The Consequences of Exposure to Inaccurate Information. *Policy Insights from the Behavioral and Brain Sciences*, 5(2):232–239, October 2018.

[92] Daniel Jolley, Mathew D. Marques, and Darel Cookson. Shining a spotlight on the dangerous consequences of conspiracy theories. *Current Opinion in Psychology*, 47:101363, October 2022.

[93] Lotte Pummerer, Robert Böhm, Lau Lilleholt, Kevin Winter, Ingo Zettler, and Kai Sassenberg. Conspiracy Theories and Their Societal Effects During the COVID-19 Pandemic. *Social Psychological and Personality Science*, 13(1):49–59, January 2022.

[94] Valerie van Mulukom, Lotte J. Pummerer, Sinan Alper, Hui Bai, Vladimíra Čavojová, Jessica Farias, Cameron S. Kay, Ljiljana B. Lazarevic, Emilio J. C. Lobato, Gaëlle Marinthe, Irena Pavela Banai, Jakub Šrol, and Iris Žeželj. Antecedents and consequences of COVID-19 conspiracy beliefs: A systematic review. *Social Science & Medicine*, 301:114912, May 2022.

[95] Cass R. Sunstein and Adrian Vermeule. Conspiracy Theories, January 2008.

[96] Calum Thornhill, Quentin Meeus, Jeroen Peperkamp, and Bettina Berendt. A Digital Nudge to Counter Confirmation Bias. *Frontiers in Big Data*, 2:11, June 2019.

[97] Timothy Neff and Victor Pickard. Funding Democracy: Public Media and Democratic Health in 33 Countries. *The International Journal of Press/Politics*, page 19401612211060255, December 2021.

[98] Christina Holtz-Bacha. The kiss of death. Public service media under right-wing populist attack. *European Journal of Communication*, 36(3):221–237, June 2021.

[99] Richard H. Thaler and Cass R. Sunstein. *Nudge: Improving Decisions About Health, Wealth, and Happiness*. Penguin Books, New York, revised & expanded edition edition, February 2009.

[100] Evan Selinger and Kyle Powys Whyte. Nudging Cannot Solve Complex Policy Problems. *European Journal of Risk Regulation*, 3(1):26–31, March 2012.

[101] Stephanie Mertens, Mario Herberz, Ulf J. J. Hahnel, and Tobias Brosch. The effectiveness of nudging: A meta-analysis of choice architecture interventions across behavioral domains. *Proceedings of the National Academy of Sciences*, 119(1):e2107346118, January 2022.

[102] Gordon Pennycook, Ziv Epstein, Mohsen Mosleh, Antonio A. Arechar, Dean Eckles, and David G. Rand. Shifting attention to accuracy can reduce misinformation online. *Nature*, 592(7855):590–595, April 2021.

[103] Gordon Pennycook and David G. Rand. Accuracy prompts are a replicable and generalizable approach for reducing the spread of misinformation. *Nature Communications*, 13(1):2333, April 2022.

[104] Gordon Pennycook and David G. Rand. Nudging Social Media toward Accuracy. *The ANNALS of the American Academy of Political and Social Science*, 700(1):152–164, March 2022.

[105] Gordon Pennycook, Jonathon McPhetres, Yunhao Zhang, Jackson G. Lu, and David G. Rand. Fighting COVID-19 Misinformation on Social Media: Experimental Evidence for a Scalable Accuracy-Nudge Intervention. *Psychological Science*, 31(7):770–780, July 2020.

[106] Jon Roozenbeek, Alexandra L. J. Freeman, and Sander van der Linden. How Accurate Are Accuracy-Nudge Interventions? A Preregistered Direct Replication of Pennycook et al. (2020). *Psychological Science*, 32(7):1169–1178, July 2021.

[107] Keith Frankish. Dual-Process and Dual-System Theories of Reasoning. *Philosophy Compass*, 5(10):914–926, 2010.

[108] Hause Lin, Gordon Pennycook, and David G. Rand. Thinking more or thinking differently? Using drift-diffusion modeling to illuminate why accuracy prompts decrease misinformation sharing. *Cognition*, 230:105312, January 2023.

[109] Henner Gimpel, Sebastian Heger, Christian Olenberger, and Lena Utz. The Effectiveness of Social Norms in Fighting Fake News on Social Media. *Journal of Management Information Systems*, 38(1):196–221, January 2021.

[110] Man-pui Sally Chan, Christopher R. Jones, Kathleen Hall Jamieson, and Dolores Albarracín. Debunking: A Meta-Analysis of the Psychological Efficacy of Messages Countering Misinformation. *Psychological Science*, 28(11):1531–1546, November 2017.

[111] Sakari Nieminen and Lauri Rapeli. Fighting Misperceptions and Doubting Journalists' Objectivity: A Review of Fact-checking Literature. *Political Studies Review*, 17(3):296–309, August 2019.

[112] Nathan Walter, Jonathan Cohen, R. Lance Holbert, and Yasmin Morag. Fact-Checking: A Meta-Analysis of What Works and for Whom. *Political Communication*, 37(3):350–375, May 2020.

[113] Nathan Walter, John J. Brooks, Camille J. Saucier, and Sapna Suresh. Evaluating the Impact of Attempts to Correct Health Misinformation on Social Media: A Meta-Analysis. *Health Communication*, 36(13):1776–1784, November 2021.

[114] Ethan Porter and Thomas J. Wood. The global effectiveness of fact-checking: Evidence from simultaneous experiments in Argentina, Nigeria, South Africa, and the United Kingdom. *Proceedings of the National Academy of Sciences*, 118(37):e2104235118, September 2021.

[115] Ullrich K. H. Ecker, Stephan Lewandowsky, John Cook, Philipp Schmid, Lisa K. Fazio, Nadia Brashier, Panayiota Kendeou, Emily K. Vraga, and Michelle A. Amazeen. The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, 1(1):13–29, January 2022.

[116] Otávio Vinhas and Marco Bastos. Fact-Checking Misinformation: Eight Notes on Consensus Reality. *Journalism Studies*, February 2022.

[117] Stephan Lewandowsky and Sander van der Linden. Countering Misinformation and Fake News Through Inoculation and Prebunking. *European Review of Social Psychology*, 32(2):348–384, July 2021.

[118] Norman C. H. Wong. "Vaccinations are Safe and Effective": Inoculating Positive HPV Vaccine Attitudes Against Antivaccination Attack Messages. *Communication Reports*, 29(3):127–138, September 2016.

[119] Sander van der Linden, Anthony Leiserowitz, Seth Rosenthal, and Edward Maibach. Inoculating the Public against Misinformation about Climate Change. *Global Challenges*, 1(2):1600008, 2017.

[120] Philipp Schmid and Cornelia Betsch. Effective strategies for rebutting science denialism in public discussions. *Nature Human Behaviour*, 3(9):931–939, September 2019.

[121] Cecilie S. Traberg, Jon Roozenbeek, and Sander van der Linden. Psychological Inoculation against Misinformation: Current Evidence and Future Directions. *The ANNALS of the American Academy of Political and Social Science*, 700(1):136–151, March 2022.

[122] Jon Roozenbeek, Sander van der Linden, and Thomas Nygren. Prebunking interventions based on "inoculation" theory can reduce susceptibility to misinformation across cultures. *Harvard Kennedy School Misinformation Review*, 1(2), February 2020.

[123] Andrew M. Guess, Michael Lerner, Benjamin Lyons, Jacob M. Montgomery, Brendan Nyhan, Jason Reifler, and Neelanjan Sircar. A digital media literacy intervention increases discernment between mainstream and false news in the United States and India. *Proceedings of the National Academy of Sciences*, 117(27):15536–15545, July 2020.

[124] Melisa Basol, Jon Roozenbeek, Manon Berriche, Fatih Uenal, William P. McClanahan, and Sander van der Linden. Towards psychological herd immunity: Cross-cultural evidence for two prebunking interventions against COVID-19 misinformation. *Big Data & Society*, 8(1):20539517211013868, January 2021.

[125] Rakoen Maertens, Jon Roozenbeek, Melisa Basol, and Sander van der Linden. Long-term effectiveness of inoculation against misinformation: Three longitudinal experiments. *Journal of Experimental Psychology: Applied*, 27:1–16, 2021.

[126] Emily K. Vraga, Leticia Bode, and Melissa Tully. The Effects of a News Literacy Video and Real-Time Corrections to Video Misinformation Related to Sunscreen and Skin Cancer. *Health Communication*, 37(13):1622–1630, November 2022.

[127] Josh Compton, Sander van der Linden, John Cook, and Melisa Basol. Inoculation theory in the post-truth era: Extant findings and new frontiers for contested science, misinformation, and conspiracy theories. *Social and Personality Psychology Compass*, 15(6):e12602, 2021.

[128] Li Qian Tay, Mark J. Hurlstone, Tim Kurz, and Ullrich K. H. Ecker. A comparison of prebunking and debunking interventions for implied ver-

sus explicit misinformation. *British Journal of Psychology*, 113(3):591–607, 2022.

[129] Nadia M. Brashier, Gordon Pennycook, Adam J. Berinsky, and David G. Rand. Timing matters when correcting fake news. *Proceedings of the National Academy of Sciences*, 118(5):e2020043118, February 2021.

[130] Michelle A Amazeen. Journalistic interventions: The structural factors affecting the global emergence of fact-checking. *Journalism*, 21(1):95–111, January 2020.

[131] Simone Chambers. Deliberative Democratic Theory. *Annual Review of Political Science*, 6(1):307–326, 2003.

[132] Joshua Cohen. Deliberative Democracy. In Shawn W. Rosenberg, editor, *Deliberation, Participation and Democracy: Can the People Govern?*, pages 219–236. Palgrave Macmillan UK, London, 2007.

[133] Jürgen Habermas. *Kommunikatives Handeln und detranszendentalisierte Vernunft*. Number 18164 in Universal-Bibliothek. P. Reclam, Stuttgart, 2001.

[134] Rishab Bailey, Vrinda Bhandari, and Faiza Rahman. Examining the Online Anonymity Debate: How Far Should the Law Go in Mandating User Identification?, June 2021.

[135] Cass R. Sunstein. The Ethics of Nudging. *Yale Journal on Regulation*, 32:413, 2015.

[136] Thomas RV Nys and Bart Engelen. Judging Nudging: Answering the Manipulation Objection. *Political Studies*, 65(1):199–214, March 2017.

[137] T. M. Wilkinson. Nudging and Manipulation. *Political Studies*, 61(2):341–355, June 2013.

[138] Henry Farrell and Cosma Rohilla Shalizi. Pursuing Cognitive Democracy. In Danielle Allen and Jennifer S. Light, editors, *From Voice to Influence: Understanding Citizenship in a Digital Age*, pages 209–231. University of Chicago Press, Chicago ; London, 1st edition edition, June 2015.

[139] Joseph E. Uscinski and Ryden W. Butler. The Epistemology of Fact Checking. *Critical Review*, 25(2):162–180, June 2013.

[140] Michelle A. Amazeen. Revisiting the Epistemology of Fact-Checking. *Critical Review*, 27(1):1–22, January 2015.

[141] Joshua A. Compton and Michael Pfau. Inoculation Theory of Resistance to Influence at Maturity: Recent Progress In Theory Development and Application and Suggestions for Future Research. *Annals of the International Communication Association*, 29(1):97–146, January 2005.

[142] David C. Logan. Known knowns, known unknowns, unknown unknowns and the propagation of scientific enquiry. *Journal of Experimental Botany*, 60(3):712–714, March 2009.

[143] Michael Patty. Social Media and Censorship: Rethinking State Action Once Again. *Mitchell Hamline Law Journal of Public Policy and Practice*, 40:99, 2019.

[144] Siu Kit Yeung, Tijen Yay, and Gilad Feldman. Action and Inaction in Moral Judgments and Decisions: Meta-Analysis of Omission Bias Omission-Commission Asymmetries. *Personality and Social Psychology Bulletin*, 48(10):1499–1515, October 2022.