Misinformation as a Negative Externality:
Theory to Remedy

Ryan Cattich

Presented to the Department of Economics
in partial fulfillment of the requirements
for a Bachelor of Arts degree with Honors

Boston College
Chestnut Hill, Massachusetts
May 2022

Abstract

In the wake of events like the COVID-19 pandemic, the storming of the Capitol, and the Russian invasion of Ukraine, it's time to start labeling misinformation for what it is: a negative externality to society. The spillover effects from the proliferation of mis- and disinformation have the potential to negatively impact the institution of democracy, civic engagement, and downstream health outcomes. Put simply, to understand the misinformation problem is to understand its complexities, its pitfalls, and its motivations. Taken as a whole, this paper articulates the need for a divergence from conventional economic theory on efficiency to a pro-social, welfare-based approach to internalization efforts. In doing so, this analysis presents a full-scale characterization of misinformation as a negative externality, starting with the reorganization of traditional microeconomic theory, followed by a platform-by-platform evaluation of various internalization strategies and evidence from the literature on the impacts of misinformation, and concluding with a commentary on potential remediation approaches.

Keywords: negative externality, misinformation, disinformation, economic efficiency, digital platforms

Acknowledgements

Above all, I owe an enormous debt of gratitude to the Economics Department for introducing me to a world of interdisciplinary research I never thought possible. Studying economics these past four years (especially econometrics) forced me to both question everything and to be patient with complexity, fueling my personal development and shaping my worldview. This thesis is truly a culmination of my undergraduate experience at Boston College; outside of Economics, my concentration in Hispanic Studies taught me about how civilizations across time, language, and geography attempt to define democracy, along with what the very institution of economics should look like and the many forms efficiency takes on. That being said, this paper is my attempt at making sense of this ultra-polarized and often hateful world we live in, focusing not only on how we got to this point, but where we can go from here.

Back in 2019, I attended the book premiere for *Good Economics for Hard Times*, and hearing directly from Abhijit Banerjee and Esther Duflo really inspired me to try to tackle one of the greatest issues facing society today, and to do so by placing human dignity at the forefront of everything. Abhijit and Esther, you may never know me, but thank you; your vision is contagious.

Such a paper would not exist if it were not for my advisor, Professor Thomas Wesner, and his persistent support throughout this process. Wes, you are truly my biggest 'hype man' and inspire the best out of me – and for that I'm forever grateful.

Special thanks is given to Gabrielle Lim and Alexei Abrahams from the Shorenstein Center on Media, Politics, and Public Policy and at the Harvard Kennedy School, your research, guidance, and advice was instrumental to this project.

Finally, thank you to my friends and family for their constant support throughout this process, and sorry for being MIA for months on end – I promise I was doing something important. Also, thank you to each and every member of the BC Bigs leadership team for your support and encouragement on this journey, I appreciate you all. Thank you especially to Molly Myman for the editing and writing support!

And Jess, thank you for encouraging me from the very start and for supporting me on a daily basis thereafter – and for being the best Bapst partner I could ask for.

Contents

1 Introduction

"In times of universal deceit, telling the truth will be a revolutionary act."

— George Orwell

In November of 2021, droves of QAnon devotees gathered in Dallas, Texas under the pretense that John F. Kennedy Jr. – who died in 1999 – would somehow return on the anniversary of his father's assassination.[1] In November of 2020, South Dakota nurse Jodi Doering recounted a few days' worth of COVID-19 patients in a Twitter thread, noting: "The ones that stick out are those who still don't believe the virus is real…They call you names and ask why you have to wear all that 'stuff' because they don't have COVID because it's not real."[2] As of February 2020, Vladimir Putin justified an invasion of Ukraine with the fabricated pretext of "de-Nazification;" de-Nazifying a country whose democratically-elected, Jewish President lost several relatives in the Holocaust (Fischer and Basu, 2022). Suffice it to say, misinformation is everywhere.[3] In fact, there is no substantial evidence to suggest George Orwell ever said or wrote the above quote, which is commonly misattributed by authors around the world. From conspiracy theorists to Kremlin propagandists, misinformation has become an unavoidable threat to the fabric of society.

This bleak reality begs the following question: do we live in a "post-truth" era? The pervasiveness of misinformation in 2016 led the Oxford English Dictionary (OED) to nominate post-truth as the word of the year, representative of circumstances

[1] See: "QAnon Supporters Pack Site of JFK Assassination." (November 2021).
[2] Doering, Jodi. Twitter Post. Nov. 14, 2020, 4:32 PM.
[3] Note: I use the prefix "mis-" intentionally here; See: "Defining misinformation and disinformation"

in which objective facts are less influential in affecting public opinion than appeals to emotion or personal belief.[4] Albeit important to highlight, the problem is less about post-truth and more about the asymmetry of information – or the media ecosystems that allow for the production of moral hazards with asymmetrical information. Put simply, there is an *oversupply* of facts in the 21st century; i.e., too many sources and too many methods, and all with varying levels of credibility. A post-truth society assumes an obfuscation of fact, whereas in reality anyone, irrespective of their echo chamber, can find anything on the internet to confirm their beliefs, biases, or bigotry. Thus, society today is – and will be – characterized by how people react, interact, and respond to the misinformation problem.

As a precondition to solving this problem, misinformation must first be acknowledged for what it is: a negative externality to society. As such, this paper functions as both a 'how did we get here' and 'where we're going,' but also as a call to action: mitigating the decentralized nature of misinformation interventions through a number of remediation strategies. As we approach four billion social network users worldwide, any analysis of the effects of misinformation *must* be centered around the digital platforms in which misinformation both originates and disseminates. Under this framework, I integrate and adapt misinformation into the existing economic theory of externalities, representing the harms of misinformation as negative externalities to society. Outside of these representations, this paper exists at the crux of law and economics: how can legislation on platform governance, and the

[4] See: "Oxford Word of the Year 2016."

subsequent interventions it allows for, provide the appropriate incentives for platforms to internalize the externalities of misinformation? In search of a remedy – outside of the traditional economist Pigouvian tax solution – it is important to begin at the root of the problem, the be-all, end-all for internet regulation: Section 230 of the Communications Decency Act.

Using the theory behind externalities and the statute of Section 230 as a foundation, I offer up substantial evidence and commentary as to how misinformation negatively impacts society in three key areas: health outcomes, civic engagement, and the democratic process. In the process, the analyses and proposals in this paper add to the literature on the effects of media and misinformation on behavior and health outcomes (La Ferrara, 2016; DellaVigna and La Ferrara, 2015; La Ferrara et al., 2012; Chiang and Knight, 2011; Jensen and Oster, 2009). Prior investigations have revealed that media exposure to misinformation can increase hate crimes (Muller and Schwarz, 2018; Bursztyn et al., 2019) and mass killings (Yanagizawa-Drott, 2014); it can also affect domestic violence (Card and Dahl, 2011), fertility choices (La Ferrara et al., 2012; Kearny and Levine, 2015), and responses to natural disasters (Long et al., 2019).

Above all, this paper exists as a top-down approach to representing misinformation as a negative externality to society – from *theory* to *remedy*. Beginning with the theory, I start by laying the necessary statutory foundation that permits digital platform interventions. In doing so, I present a brief history of Section 230 and detail its subsequent clauses to highlight the present-day ambiguity in the statute. From there, I build off of the conventional microeconomic theory on how to represent misinformation as a negative externality, while offering some key definitions and qualifications for mis- and disinformation. Next, I present platform-level analyses on the unilateral internalization strategies among the most prominent digital platforms, evaluating how these platforms operate within the ambiguous scope of Section 230 protections. To conclude, I assess a non-exhaustive list of potential remediation strategies for the externality problem, ranging from technocratic solutions to thought experiments.

2 Statutory Context: Section 230

2.1 The Communications Decency Act (CDA)

Introduced in February of 1995 by Senator James Exon, the Communications Decency Act (CDA) was initially created to combat a growing issue of extensive pornography and obscenity on the internet (Cannon, 1996). As passed, the CDA extends the antiharassment, indecency, and antiobscenity restrictions currently placed on telephones to interactive computer services, or ICPs – they will be referred to herein as "digital platforms" for the sake of simplicity.[5] The bill was promptly met with resistance from some lawmakers and interest groups who opposed the idea of meddling with the internet, calling the bill a violation of free speech. Opposing the CDA based on these concerns, coupled with the 1995 New York Supreme Court decision *Stratton Oakmont v. Prodigy Services Co.*, representatives Christopher Cox and Ron Wyden proposed the "Cox-Wyden Amendment" to Exon's bill, some parts of which would ultimately become Section 230 of the CDA as it stands today.

In *Stratton Oakmont*, the court held that a digital platform could be held liable for defamatory content posted by users on its platform, given that the platform *proactively* monitored, screened, and removed offensive user content; thus, the platform serves as an editor *and* publisher of all posted content thereby assuming legal responsibility (*Stratton Oakmont v. Prodigy Servs. Co.*, 1995). By contrast, a

[5] Note: Legal literature uses "ICSP" to characterize interactive computer services providers, while other scholars default to "ISPs." "Digital platforms" is the most straightforward nomenclature, yet the connotations should not be oversimplified.

1991 New York case, *Cubby, Inc. v. CompuServe, Inc.*, held that a digital platform that did *not* regulate third-party user content avoided liability for libel since it did not know of and had no editorial control over posted defamatory material. Shockingly, the court in *Stratton Oakmont* asserted that a platform that does not perform any intervention mechanisms for 'problematic' content can never be legally responsible for the content of its users. On the other hand, a service that takes voluntary, *bona fide* action to screen such content subjects itself to liability (Ardia, 2010).

As a direct response, the Cox-Wyden Amendment was proposed to incentivize digital platforms to take proactive measures to improve online safety and regulate objectionable content – without the fear of liability. This amendment allowed private platforms to address the problem of online indecency, while simultaneously upholding the Representatives' policy goal of fostering the "vibrant and competitive free market" that is the internet (47 U.S.C. § 230). Shortly after, Congress passed the Telecommunications Act of 1996, which included the CDA and the Cox-Wyden Amendment – legislation presently regarded as Section 230.

2.2 What is Section 230?

Section 230 of the CDA is the most consequential piece of legislation regarding internet regulation as we know it. Many – if not all – of the prominent digital platforms depend on Section 230; given the prevalence of social media today, around four billion of us rely on Section 230-enabled services daily. From its inception in 1996, Section 230 laid the foundation for the 'Big Tech' conglomerates of the 21st century, pledging liability protection to companies for all third-party content posted and shared on their platforms.

Per the statute, such protection from liability is extended to all ICPs, denoted as any "information service, system, or access software provider that provides or enables computer access by multiple users to a computer server;" i.e., any platform that provides access to the internet.[6] The law defines all users of digital platforms as "information content providers," thus assigning responsibility to any person or entity who creates or develops information provided through the internet, in whole *or* in part. By proxy, Section 230 covers the vast majority of websites and internet-based applications, everywhere from Google to 8chan and every user in between. Given this expansive reach, subsection (b) outlines the policy positions of the United States:

> **1.** to promote the continued development of the Internet and other interactive computer services and other interactive media;
> **2.** to preserve the vibrant and competitive free market that presently exists for the Internet and other interactive computer services, unfettered by Federal or State regulation;
> **3.** to encourage the development of technologies which maximize user control over what information is received by individuals, families, and schools who use the Internet and other interactive computer services;
> **4.** to remove disincentives for the development and utilization of blocking and filtering technologies that empower parents to restrict their children's access to objectionable or inappropriate online material; and
> **5.** to ensure vigorous enforcement of Federal criminal laws to deter and punish trafficking in obscenity, stalking, and harassment by means of a computer.[7]

[6] See: U.S.C. § 230 (f)(2).

[7] See: U.S.C. § 230 (b).

Bearing these objectives in mind, the primary aim of Section 230 was to safeguard the public interest in monitoring, blocking, and screening objectionable content on the internet, but to do so in such a way that enabled and incentivized digital platforms to grow without fears of onerous regulation and endless litigation. To qualify for liability immunity, digital platforms must under the 'Good Samaritan' blocking and screening guidelines as set in the statute; thus, subsection (c) represents the most consequential legislation of Section 230:

> **1. Treatment of Publisher or Speaker**
> No Provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider.
>
> **2. Civil Liability**
> No provider or user of an interactive computer service shall be held liable on account of:
>
> > **A.** any action voluntarily taken in good faith to restrict access to or availability of material that the provider or user considers to be obscene, lewd, lascivious, filthy, excessively violent, harassing or otherwise objectionable, whether or not such material is constitutionally protected; or
> > **B.** any action taken to enable or make available to information content providers or others the technical means to restrict access to material described in paragraph **1**.[8]

With these statutes, Section 230 enabled digital platforms, like Facebook and Twitter, to grow and thrive without crippling legal exposure or expensive editorial staffs; yet, today, Section 230 casts a foreboding shadow over the internet and, by extension, society at large. The digital landscape of the internet has surely evolved

[8] See: U.S.C. § 230 (c).

since 1996, yet Section 230 has not. In the crosshairs of a novel pandemic and authoritarian aggression in the 21st century, exceptions to liability protection have never been more important; therefore, Section 230 needs more specificity when it comes to when platforms do not have liability immunity. As written, Section 230 does not permit liability immunity for some specific offenses: violations of federal criminal law, communications privacy law, intellectual property law, and federal sex trafficking law. However, Section 230 is currently lacking specificity with respect to what explicit actions separate platforms from publishers, what constitutes such good faith voluntary action, and how these ambiguous policy mechanisms can be *efficiently* achieved.[9] Today, courts are progressively interpreting Section 230 to broadly restrict the scope of civil and state liability for digital platforms, sheltering web-based corporations against a multitude of claims (Goldman, 2017). Yet, in the absence of specificity in subsection (c), how digital platforms moderate within the voluntary, *bona fide* framework is, evidently, up to them.

In addition to reaching key objectives, the clauses of Section 230 are all about incentives. Thus, Section 230 is at the crossroads of law and economics, in that its policy mechanisms have the potential to incentivize both users and platforms to reach efficient outcomes; e.g., outcomes where groups or individuals do not suffer the intended or unintended consequences of misinformation. With the categorization of misinformation as a negative externality, Section 230 must intentionally encourage

[9] Note: Here, efficiency is achieved upon completion of the core objectives: incentivizing further development of the internet, preserving the free market of ideas, allowing for innovation concurrent with user control, and doing so in a secure, iterative way. But, in a way least reminiscent of George Orwell's *1984* "Big Brother," yet still prevents disinformation from becoming a negative externality.

risk prevention on a platform-by-platform basis, within specific conditions for what does or does not warrant an exception to liability protection.[10] To better understand the remedies, it is essential to first lay the groundwork for the theory of misinformation as a negative externality.

[10] See: The 'Remedies' section for statutory modifications.

3 Theory and Foundation

Over one hundred years ago, prominent British economist Arthur Pigou first identified the problem of externalities: a business that could not absorb all of the costs associated with the goods it produced and sold. Found in almost every 'Econ 101' textbook, some classic examples of negative externalities include the effects of pollution on the environment and secondhand smoke on humans. Today, in addition to the carcinogenic effects of chemical runoffs and tobacco smoke, we have to contend with a new problem: the disintegration of democracy, intentional disinformation campaigns, and incitements to violence involuntarily enabled by Facebook, YouTube, and other major digital platforms (Verveer, 2019). Taking it one step further, misinformation is an invisible pollutant, worthy of attention from economists and not isolated to the private sector – misinformation concerns and affects everyone.

3.1 Defining misinformation and disinformation

"Disinformation is a broad category that's tough to define, and resulting liability for its harms is very uncertain."

— Professor Matthew Waxman [11]

Here, Matthew Waxman, a professor at Columbia Law School and specialist on national security, is both right and wrong. Right in the sense that 'disinformation' is difficult to classify, but wrong in the sense that the definitions are *clear* and the harms are apparent. Information scientists have long contemplated the nature of

[11] See: "FCC seeks input on regulatory review of 'Section 230' liability protection; issues touch on countering disinformation," *Inside Cybersecurity*, 2020.

information: how to define it, where it comes from, the kind of actions it affords individuals, and so on. Yet in particular, mis- and disinformation tend to be limited and misunderstood areas in the effort to understand the very *nature* of information; evidenced by news cycles and commentators mistakenly using mis- and disinformation interchangeably, conflating their definitions (Rubin, 2010). In short, the main difference between misinformation and disinformation is *intent*; misinformation should be used when the intent to mislead is unproven, whereas disinformation should be used when the intent to mislead is well-defined; for this reason, 'propaganda' is commonly associated with disinformation.

Yet the very nature of information is such that information – and subsequently mis- and disinformation – is relative. According to Tuominen and Savolainen (1997), vis-à-vis a *social constructionist view*, the nature of information can be best represented as a "communicative construct which is produced in a social context." Such a constructionist view of information is valuable when discussing mis- and disinformation because it emphasizes a people-centric approach to social context, using conversations between people as ways of determining what information is and what can be informative. Mis- and disinforming are information *actions* that occur in discourse between people, both face-to-face and on digital platforms. So, through any channel of communication, mis- and disinformation can be information people use to "construct some reality" (Karlova and Fisher, 2013); in this way, mis- and disinformation can be characterized as *extensions* of information. Irrespective of context, though, the definitions are clear: misinformation is unintentionally incorrect information and disinformation is systematic and intentional deception.

But given the circumstantial nature of information, the scale of the misinformation problem is highly contested and contingent upon *how* misinformation is classified (Rogers, 2020). While the OED defines misinformation as, "wrong or misleading information," relatively little elaboration exists in the current literature; authors generally cite the OED definition without further commentary or exploration: Bednar and Welch (2008) and Stahl (2006), to name a few. On the other side of the equation, Fox (1983) makes the argument that "information need not be true," implying that there is no reason information *must* be true so misinformation, transitively, may be false. Fox states that "misinformation is a *species* of information," outlining the relationship between misinformation and information as follows: misinformation, albeit false, is still information by definition and, therefore, can still be informative. With respect to ambiguity, Fox is correct in this philosophy. Misinformation, given its indefinite, yet persistent nature on digital platforms, does inform people – often at their expense. Misinformation, given the uncertainty of its intent, takes on countless forms: inaccuracy, uncertainty, unclear phrasing or general vagueness, and being open to multiple interpretations (Karlova and Lee, 2011). However, incomplete information may also qualify as a form of deceit, which would then be classified as disinformation.

According to the OED, disinformation is designated as "deliberately false information," especially when supplied by a government to influence the opinions and policies of those on the receiving end. Ironic enough, the OED notes the term derives from Russian, translated as *dezinformacija*, and first recorded in 1949. Given the political and cultural panoramas in the Soviet Union at the time, the underlying connotations between disinformation and negative, malicious intent most likely developed as a result of Stalinist information control policies at the time. Yet according to a 2015 European Parliament briefing, Russia listed disinformation as one of the main threats to international peace and security, defining the term as the manipulation of the flow of information with the intent of adversely affecting the "psychological or spiritual state of society, or eroding traditional cultural, moral, ethical, and aesthetic values." So, per Russian criteria, the unsubstantiated claims of seeking to de-Nazify Ukraine as a basis for war qualify as intentional pointed disinformation to uproot the fabric of Ukrainian society.

Still, disinformation should never be considered as a subset of misinformation. While disinformation may share qualifications with information and misinformation, disinformation is distinctively deliberate – even though the intentions behind such deception may not always be well-defined. For instance, intentions behind disinformation could be socially-motivated or benevolent; e.g., lying to conceal a surprise party or lying to an interviewer about job qualifications. More consequentially, intentions could also be personally-motivated and antagonistic; e.g., controlling a populace or overturning a democratically-held election. Therefore, given misinformation may be false, and disinformation may be true, misinformation and disinformation must be regarded as distinct subgroups of information.[12]

For this analysis, misinformation is used as the more general term, but any use of misinformation over disinformation, and vice versa, is intentional. Given these denotational specifications, we can then move to build up and correlate the existing theory behind externalities with misinformation.

3.2 Theory behind externalities

"*Restoring human dignity to its central place…sets off a profound rethinking of economic priorities and the ways in which societies care for their members*."

– Abhijit V. Banerjee and Esther Duflo [13]

The crucial feature of externalities is that there are things people care about that cannot be sold in a market: there is no market for loud music in the middle of the night, wandering smoke on a casino floor – or misinformation. It is this lack of markets for externalities that causes problems; their absence implies poor, if any, regulation, negative economic consequences, and asymmetrical information. To restore human dignity, as Banerjee and Duflo suggest, this necessitates a reconsideration of priorities around how we define and qualify efficiency.

The conventional economic theory behind externalities concerns two main functions: consumption and production. A consumption externality arises when one consumer cares directly about another's production or consumption, and a production

[12] See: Appendix Table 1.

[13] See: *Good Economics for Hard Times* (2019).

externality when the production possibilities of one firm are influenced by consumers or another firm. Let's suppose an interaction between a Toyota dealership in Watertown, Massachusetts, and local consumers, where the dealership's end goal is profit maximization. COVID-19 supply chain bottlenecks on the West Coast – another unfortunate negative externality – present unforeseen obstacles to the dealership's supply, forcing them to make constrained production decisions. In other words, it's on the Toyota dealership to either incur the cost of the delays, or subject Eastern Massachusetts buyers to higher prices. According to the theory behind externalities, private sector solutions to internalizing such externalities can result in quantity and price optima, just as long as one party is assigned *property rights*.

This emphasis on the distribution of property rights is often regarded as the Coase Theorem, coined by economist Ronald Coase. According to the Coase Theorem, the practical problems with externalities generally arise as a result of poorly defined property rights (Varian, 2014). Such is the ambiguous case with misinformation – can we assign property rights to platforms or platform users? Who is the responsible party to be assigned property rights for disseminating misinformation? When property rights are identifiable and well-defined, the Coase Theorem suggests property rights can bring about a *socially optimal market quantity*. To achieve this, thereby internalizing the externality, Coase calls for a reassignment of property rights. Although, in practice, such Coasian theory is likely to be more effective for small, localized externalities than for larger, boundless externalities like misinformation. Thus, when classifying misinformation as a *public sector* externality, challenges to the Coase Theorem arise. Can we arbitrarily set the socially optimal market quantity for misinformation? Who, or what institution, can determine the assignment of property rights? What would a reassignment of property rights look like? In essence, the traditional economist answer of defaulting to market mechanisms to reach points of efficiency does not work in the case of misinformation.

On the topic of efficiency, especially in the context of externalities, theoretical economists generally emphasize the importance of *Pareto efficient*, or *Pareto optimal* solutions (Sandler and Smith, 1976; Luenberger, 1992). In sum, a Pareto efficient situation is one where no individual can be made better off without making someone else worse off; or the condition where we cannot improve the utility of any individual without decreasing the utility of another individual (Varian, 2014). Thus, in the case of misinformation filtering, a Pareto efficient condition would arise when we seek to maximize social welfare – or minimize social costs – at the *optimum* level of misinformation. But what is the optimum level of misinformation?[14]

Such assumptions create complications for both theory and policy for filtering, yet, with respect to internalizing the externalities of misinformation, Pareto efficiency is *not* as important. An important qualification of Pareto efficiency is that it does not necessarily imply equitable outcomes, i.e., a society with Pareto improvements can still have inequity within it (Sanders, 2021). Moreover, Pareto efficiency acts as an inhibitor to intervention by a third party; e.g., state legislatures or Congress. With any degree of assumed perfect competition in a market, any

[14] Note: While 0 may be an obvious numerical optimum, this assumption could raise concerns about the limitation of free speech.

equilibrium will then tend to be Pareto optimal, meaning a Pareto improvement is not possible by intervening; effectually, any intervention that will make one individual better off will make someone else worse off. All of this is not to say that efficiency is not important, or not achievable. Rather, the measure of efficiency must be redefined to account for internalizing misinformation externalities.

More applicable in practice and with less stringent criteria, a better form of economic efficiency to consider is Kaldor-Hicks efficiency. Under the criterion for Kaldor-Hicks improvement, an outcome is an improvement if the benefiting party could theoretically 'compensate' the party that loses out. Or, if a change adds more to the *happiness* of the individuals that benefit than it does the *sadness* of those who lose out – in the interest of the common good (Sanders, 2021). Under this quasi-utilitarian framework, we can more broadly classify and evaluate any potential improvements to the misinformation internalization problem, in addition to informing policy decisions concerning the effectiveness of Section 230. For misinformation interventions, the principal objective should be to minimize social costs and maximize social benefits. Thus, when assessing the evidence of misinformation intervention mechanisms, Kaldor-Hicks improvements imply a multitude of cost-benefit analyses, where the benefits of internalizing misinformation should outweigh the costs. In the context of misinformation, however, such cost-benefit analyses should be more focused on both the implied and actual *social* costs, rather than numerical costs.

With respect to the Kaldor-Hicks framework, this implies the compensation component is of no importance to the misinformation problem. For example, consider pollution as the externality: a voluntary exchange between two pollution-generating parties qualifies as a Kaldor-Hicks improvement if the buyers and sellers would still carry out the transaction even if it means fully compensating the victims of the pollution. While this equilibrium signals a more efficient market, it does not solve the pollution problem, given that the firms are still allowed to pollute freely. Given there are no buyers and sellers of misinformation – at least formally – the compensation component, albeit largely theoretical, assumes financial transactions can function as sufficient intervention mechanisms. Substituting misinformation for pollution in this example, it is clear that Facebook cannot practically compensate the Rohingya people persecuted in Myanmar; thus, going forward, it is clear the compensation component of the Kaldor-Hicks framework need not be considered further. Outside of compensating the affected parties of pollution, economists have come up with a multitude of other solutions to the pollution problem; but how do some of these solutions stack up to the misinformation problem?

3.3 Misinformation as pollution: a thought experiment

Metaphorically, misinformation is the pollution of information. More pointedly, misinformation is to Twitter what pollution is to the environment – an intangible, invisible pollutant to society. While several economic models attempt to solve the pollution internalization problem between firms, this is not the case for misinformation. Given the symbolic nature of the pollution-misinformation relationship, it may be advantageous to reconfigure a traditional economic model of corrective measures to better inform approaches to internalizing the externalities of misinformation. Thus, let's construct a model of emissions regulation and compare it with misinformation regulation between two firms: Models 1 and 2, respectively.

For this thought eperiment, *marginal damage* (MD) represents any additional costs associated with the production of a good or service that is imposed on others but that producers do not pay (Varian, 2014); therefore, MD is assumed to be ≤ 0.[15] *Private marginal cost* (PMC) represents the direct cost to producers of producing an additional unit of a good or service, and *social marginal cost* is defined as the private marginal cost to producers plus marginal damage; for example, a steel plant that pollutes a river but does not face any pollution regulation quotas will inevitably ignore pollution altogether when deciding how much to emit. *Social marginal cost* (SMC), therefore, is represented by this simple equation $SMC = PMC + MD$. Given Kaldor-Hicks motivations for achieving efficiency, minimizing social costs is paramount for both pollution and misinformation quantity regulation.

As motivation for Model 1, consider the case of carbon dioxide emissions. One firm may find it relatively inexpensive to reduce its emissions of CO_2, whereas another may find it more expensive. Given this, Model 1 carries the following assumptions: there are only two firms, Firm 1 and Firm 2; Firm 1's emission quota is x_1 and Firm 2's is x_2; the cost of achieving emission quotas is $c_1(x_1)$ for Firm 1, $c_2(x_2)$ for Firm 2; and the total amount of emission is fixed at some target level, X. So, to

[15] Note: In practice, marginal damage does not always have a monetary value ascribed. Such is the case for misinformation, where resulting costs are not always quantifiable.

minimize the total costs of achieving the emissions target, subject to aggregate restraint, we must solve the following problem:

$$\min_{x1,\, x2} c_1(x_1) + c_2(x_2) \tag{3.1}$$
$$\text{such that } x_1 + x_2 = X$$

Varian (2014) suggests that a traditional economic argument implies that the marginal cost of emission control must be equalized across both firms. If one firm had a higher marginal cost of emission control than the other firm, then it would be possible to lower total costs by reducing its quota and thereby increasing the quota of the other firm; but how is this outcome attained? If government regulators held asymmetrical information on the cost of emissions for all firms, they could then estimate the appropriate arrangement of production and impose it on all relevant parties. However, according to Varian (2014), the cost of obtaining information on such costs and keeping it up-to-date is staggering; in other words, it is much easier to characterize the 'optimal solution' than to implement it.

Many economists have argued that the best way to implement an efficient solution to the emission control problem is to use a market (Stevenson, 1992), yet recall that one of the pillars of externalities is that they do not exist as markets. To sidestep this, we can create a theoretical market; with pollution, for example, such a market would be a carbon cap-and-trade system. In this system, the largest polluters in a given region are all assigned a quota for their emissions of CO_2. If the firm exactly meets its emissions quota, it faces no fines or penalties, but if a firm reduces its emissions by more than the target quota, it can sell the extra "right to emit" on the open market. Under this system of carbon credits, each firm can compare the market price of an emission credit to the cost of reducing its emissions and decide whether it was more cost-effective to reduce emissions further or purchase emission credits from other firms. Firms that easily reduce emissions will sell credits to firms that find it costly to reduce emissions. In equilibrium, the market price of the right to emit one ton of pollution should equal the marginal cost of reducing emissions by one ton. Thus, this presents an *optimal* pattern of emissions that minimizes PMC, i.e., a Kaldor-Hicks optimum for producers of emissions, but not a Kaldor-Hicks improvement for those who bear the MD of pollution. In sum, the market-based solution to the emissions control problem does little to minimize SMC, resulting in an inefficient outcome and still allowing for the emission of CO_2, albeit at potentially lower levels. Equation 3.1 rests on the assumption that there is a set target of emissions X, which allows the firms to minimize their pollution quotas with respect to one another; but in the case of misinformation, this assumption is not possible. So, what would such a market-based solution look like?

As motivation for Model 2, suppose misinformation interventions act as heterogeneous costs to platforms.[16] Thus, Model 2 carries the following assumptions: there are only two platforms, Google (G) and Twitter (T); each platform has either a low (L) or high (H) cost of misinformation reduction or filtering (r). As such, the main outcome variable is exposure – the exposure of mis- and disinformation to platform users – and exposure is quantified by the costs to reduce exposure. Twitter acts as

[16] Note: Examples of costs to platforms for regulation include, but are not limited to legal fees and frontend infrastructure.

the firm with low costs to *r*, and Google as the high-cost firm. The discrepancies in cost to *r* between the two platforms are demonstrated by the following equations:

$$G_H(r) = 1.5r^2 \Rightarrow MC_H(r) = c'_H\,(r) = 3r \qquad (3.2)$$

$$T_L(r) = 0.75r^2 \Rightarrow MC_L(r) = c'_L\,(r) = 1.5r \qquad (3.3)$$

As illustrated, it costs Google twice what it costs Twitter to perform misinformation interventions. In the absence of interventions like taxes or regulations – or when facing ambiguity as to what *bona fide* actions to pursue – firms like Twitter, Google, and the like are not incentivized to perform interventions. Thus, we can assume that with no interventions, firms set $r_L = r_H = 0$; *MD* is arbitrarily $1 per unit of misinformation for simplicity's sake. Given this, social welfare maximization can be represented as a *social marginal benefit* (SMB):

$$SMB = \max_{rH,\, rL} r^H + r^L - G_H(r^H) \;-\; T_L(r^L) \qquad (3.4)$$

$$\Rightarrow MD_H = 1,\ MD_L = 1 \Rightarrow r^H = 1/3,\ r^L = 2/3$$

Assumptions aside, this model illustrates a market optimization in which social welfare is maximized when firms that have a low cost to perform interventions (*r*) perform more interventions. In other words, an unrealistic, but theoretical Kaldor-Hicks optimum is to have the firm with the lower cost to regulation (Twitter) perform more misinformation regulation or filtering than the high cost to *r* firm (Google). The decentralized nature of this problem has cumulative effects; unilateral policy decisions partition the equilibrium in which wider platforms perform significantly more moderation than smaller, fringe platforms. This produces several, fragmented media ecosystems where generalized optima are difficult to achieve; if Twitter successfully performs a lot of moderation, who is to say that users cannot simply migrate to another platform?

The above allocation functions more or less like a carbon cap-and-trade system, yet this market mechanism does not hold the same assumptions as of the market for emissions. By nature, Twitter, even if it has the lowest relative costs to filter and regulate misinformation, can only operate within its platform, i.e., Twitter cannot, as a third party, regulate misinformation on any Google platforms, and platforms cannot buy or trade credits for misinformation filtering. Thus, in practice, defaulting to the market would *not* produce a socially optimal outcome for misinformation. To achieve Kaldor-Hicks efficiency with misinformation, optima must be rooted in the maximization of social benefit over the minimization of production costs, within the scope of individual platforms.

There are several useful interpretations of the conditions for Kaldor-Hicks efficiency as established above – for both pollution and misinformation. Let's suppose the two firms in Model 1 are an iron mill and a fishery, where water pollution is the negative externality; an important interpretation of efficiency here is that one firm may face the "wrong" price for pollution. As far as the iron mill is concerned, its emission of water pollution is of no cost to them; however, this neglects the costs that pollution imposes on the fishery. According to this perception, a Kaldor-Hicks improvement is achieved by making sure that the polluter (the iron mill) internalizes the appropriate social costs of its actions. One way to accomplish this is to levy a tax on the pollution generated by the iron mill, i.e., placing a tax of t dollars per unit of pollution generated by the iron mill. Such a mechanism is known as *Pigouvian taxation* – a tax on any market that generates negative externalities.[17]

From here, this becomes more of a profit maximization problem, given that the objective of profit maximization itself should encourage the internalization of the iron mill's production externalities. The central problem with Pigouvian taxes is that the optimal level of pollution must be well-defined to impose a tax. But if the optimal level of pollution could be determined, we could just tell the iron mill that exact amount. Given the qualifications and definitions of mis- and disinformation, much ambiguity surrounds what constitutes an 'optimal' level of misinformation circulation – are these theoretical optima equalized across different digital platforms, or are they heterogeneous? As such, the internalization of misinformation is unique in that misinformation comes from third-party producers on platforms, as opposed to pollution as the direct result of firms' production decisions. Moreover, misinformation internalization is *not* bounded by profit maximization, further emphasizing Kaldor-Hicks efficiency based on maximizing social welfare. In the absence of profit-maximizing motivations, this thought experiment serves as a baseline for how to inform policy mechanisms – like Section 230 and individualized community guidelines – to best structure incentives for misinformation regulation or filtering on digital platforms. However, given the ambiguity of the statute at the time of writing, platforms are left to endogenously interpret Section 230, resulting in various approaches to the misinformation internalization problem.

[17] Note: Arthur Pigou, a Cambridge University economist, explored such taxes in his book *The Economics of Welfare* in 1920; See: Section 6.1 on Pigouvian taxation.

4 Digital Platforms: Approaches to Internalization

"In a policy world that has mostly abandoned reason, if we do not intervene, we risk becoming irrelevant."

– Abhijit V. Banerjee and Esther Duflo [18]

Digital platforms of the 21st century are playing an ever-expanding role in shaping the continuously evolving information ecosystem. Events such as the COVID-19 pandemic, the 2020 U.S. election, and the Russian assault on Ukraine highlight the urgent need for decisive action, as platforms attempt to rise to the challenge of internalizing the externalities of misinformation unilaterally. Such efforts should prioritize the end goal of a well-informed – not misinformed – citizenry about matters affecting the future of democratic institutions, civic engagement, and the health outcomes of entire populations.

Individual platforms are, in essence, tasked with not only aligning their services with the needs of their users but promoting the principles of the common good as well (Cattich, 2020). In the absence of a Section 230 amendment, consequently, platforms are left to self-regulate, experimenting with and building out various regulation strategies internally. Recent proposals to amend Section 230 have progressed, as Verveer (2019) puts it, from the "artisanal to the industrial," indicating that the supply has flooded the "market." Such developments echo an undeniable reality: the business models of the largest digital platforms, for all the good they generate, enable and empower remarkably harmful activities. Thus, when evaluating

[18] See: *Good Economics for Hard Times* (2019).

the various approaches to misinformation intervention mechanisms, a platform-by-platform breakdown sheds light on which interventions best align with maximizing *social marginal benefits* and upholding Section 230's original policy goals. The scope of this analysis is confined to the following four key intervention mechanisms:

Table 1: Four main misinformation intervention mechanisms[19]

Type	Definition
Credibility label	Labels with attachments to authoritative sources or third-party fact-checkers.
Contextual label	Information that provides additional context that the content of the user-generated post does not provide.
Removal	The temporary or permanent removal of a post from a platform feed.
Downranking	Reducing the number of times a post appears in other users' social media feeds.

Given the aforementioned definitions and qualifications of mis- and disinformation, not every intervention is equally consequential. *Removal*, for example, blocks access entirely – thereby raising free speech concerns – whereas interventions like *downranking* simply reduce the distribution and frequency of content on a platform. Other, more lenient interventions such as *credibility* and *contextual labeling* allow for expanded free speech and wider distribution, yet may still provoke resistance from users (Saltz et al., 2021). To weigh the effectiveness of these various policies, the Kaldor-Hicks framework for efficiency comes into play: relying on both the platform ecosystem and circumstantial evidence to make cross-

[19] See: Appendix Table 2 for a detailed summary of the intervention mechanisms.

platform comparisons. With this foundation, we can then evaluate the effects of intervention mechanisms with respect to their contributions, or the lack thereof, to overall social benefit; i.e., the maxim that 'winners' win more than the 'losers' lose.

Regarding the implications for intervention design, Saltz et al. (2021), in evaluating self-reported attitudes toward misinformation interventions, outline four key findings for platforms to consider: *explainability*, *transparency*, *oversight*, and *trust*. In other words, respectively: making intervention sources and processes more explainable to audiences; motivating design changes with intended and actual intervention effects, acknowledging that interventions are not homogenous; considering large-scale transformations to how platforms function and relate to the public, such as external oversight; and striving to minimize errors of automated systems that reduce trust in interventions while expanding upon positive encounters (Saltz et al., 2021). These implications are paramount for platforms to consider, providing iterative benchmarks and metrics for evaluating the overall efficiency of their misinformation mitigation strategies. In tandem with these benchmarks, the key outcome variable to consider across platforms is *exposure* – i.e., what various intervention mechanisms do to limit individual users' exposure to potentially misleading or harmful information.

4.1 Meta: Facebook and Instagram

"I just believe strongly that Facebook shouldn't be the arbiter of truth of everything that people say online."

– Mark Zuckerberg [20]

4.1.A Facebook [21]

Five months after making this statement – in testimony before the United States Senate Committee on the Judiciary – then-Facebook CEO Mark Zuckerberg verbalized the platform's mission as "[giving] people the power to build community and [to] bring the world closer together," reverberating a hands-off approach to misinformation intervention (Zuckerberg, 2020). As of February 2022, however, Facebook, via their digital "Transparency Center," explicitly affirms that the platform is committed to stopping the spread of misinformation, employing a combination of "enforcement technology, human review and independent fact-checkers" as an approach to internalization (Meta, 2022).

Allegedly in practice since 2016, Facebook specifically adopts a "remove, reduce, inform" strategy for internalizing misinformation (Facebook, 2022). Emphasizing the importance of free expression, Facebook will *remove* misinformation in 3 limited cases, when misinformation has the potential to cause imminent physical harm, interfere with or suppress voting, and mislead an ordinary person via an AI-

[20] See: "Zuckerberg knocks Twitter for fact-checking Trump" (17 May 2020).

[21] Note: Given Meta is the parent company, there is an inherent overlap in policy between Facebook and Instagram; however, any distinctions between the two platforms are intentional. Despite analogous interventions, Facebook and Instagram are entirely different platforms worthy of individual evaluations. Any use of 'Meta' implicated both Facebook and Instagram with respect to policy.

manipulated video – i.e., a 'deep fake' – that replaces or superimposes content onto them (Facebook, 2022). If misinformation does not violate the Meta community standards, but may still be problematic or otherwise questionable, Facebook will *reduce* the content's distribution (Meta, 2022). For example, regarding content that some individuals may want to see, but others may find problematic, Facebook makes it more difficult to view; however, exactly how Facebook does this is not explicit enough. Lastly, as depicted in Appendix Figure 1, Facebook will *inform* users by applying labels to fact-checked posts, allowing users to view fact-checker conclusions and decide for themselves what to read, trust, or share (Facebook, 2022). These intervention mechanisms can be summarized by the following table:

	Non-harmful	**Harmful**
Confirmed misinformation	Label/Downranking	Removal
Disputed accuracy	Label	Removal
Unverified accuracy	Label	Label

With the influx of medical misinformation, the COVID-19 pandemic put Facebook's remove, reduce, inform strategy to the test: removing content that is considered to be imminently harmful and reducing or informing content that is considered not to be imminently harmful. Under this approach, Facebook removed millions of false or misleading posts related to COVID-19 and labeled other, less harmful posts with “strong warnings” (Nunziato, 2020). Facebook has also made available its Coronavirus Information Center at the top of its news feeds, a repository of curated, expert-backed information about COVID-19, alongside broadening its work with independent fact-checking organizations (Facebook, 2020). Aligning most with the *downranking* intervention, Facebook’s approach to medical misinformation revolves around reducing its influence in feed generation algorithms, and less so on removal. To the everyday user, this approach takes the form of *contextual* labels on posts that are identified as false or misleading. This resulted in Facebook issuing 40 million such warnings in March 2020 and 50 million in April 2020, with this practice, Facebook notes that 95 percent of the time users did not go on to view the original content (Rosen, 2020). Thus, on the one hand, Facebook’s overall approach to medical misinformation clearly emphasizes explainability, transparency, and trust; however, failures in the timely implementation of its interventions are problematic.

In practice, Facebook’s interventions are too reactionary. A comprehensive study was undertaken by the human rights group Avaaz examined the dissemination of over 100 pieces of misinformation about COVID-19, as indicated to be misleading, false, or harm-inducing by independent fact-checkers. Avaaz (2020) concluded that, despite intervention policies, millions of the platform’s users are continuously being put at risk, finding that pieces of misleading content were shared over 117 million times on the platform. For example, misinformation claiming that one way to rid the body of COVID-19 is to gargle water, salt, or vinegar was shared over 31,000 times before removal (Avaaz, 2020). Going further, Avaaz (2020) notes that it can take up to 22 days for Facebook to *downrank* medical misinformation and issue *credibility* or *contextual labels*; in the case of non-English content, the lag is even more severe, where 51 percent of non-English circulating misinformation had no warning labels at all. Overall, delays in the platform's execution of misinformation interventions present significant obstacles and inhibit their overall efficiency; for Facebook, therefore, the path to efficiency demands a restructuring of incentives.

In March of 2018, a U.N. fact-finding operation highlighted the role of digital platforms – Facebook, in particular – in fueling the dissemination of hate speech and disinformation against the Rohingya minority in Myanmar. The Human Rights Council (2018) determined that the violence against the Rohingya constituted genocide and that Facebook had played a "determining role" in the violence. In November of 2021, revelations from the Facebook Papers reaffirmed that Facebook's algorithmic magnification of "incendiary material," combined with the failure to prioritize moderation – especially outside of the U.S. and Europe – has ignited the spread of hate speech and misinformation.[22] Yet Facebook is not solely amplifying misinformation – the company is also funding it. According to an MIT Technology Review investigation (2021), Facebook has been paying "millions of ad dollars to bankroll clickbait actors," driving the deterioration of global information ecosystems altogether. Hao (2021) notes that "clickbait farms" have taken advantage of Facebook's lack of quality control of their "Instant Articles" program, which launched in 2015 as a way to open articles in-house – as opposed to a browser – to seize ad

[22] Note: The "Facebook Papers" is a collection of internal documents and a consortium of news organizations that were provided to congress by whistleblower Frances Haugen.

revenue from Google.[23] If a participating publisher opted into monetizing with Facebook's advertising network, Facebook could then insert ads into the publisher's stories and take 30 percent of the revenue (Hao, 2021). With negligible moderation, clickbait farms used this loophole to publish widely plagiarized content on fake websites and target dozens of Facebook pages at a time. In the case of Myanmar, clickbait actors found the right blend of engaging and provocative content, which, according to Hao (2021), could generate "thousands of US dollars a month in ad revenue," or "10 times the average monthly salary," paid to them directly by Facebook.

Suffice it to say, Facebook's policies on misinformation, albeit a step in the right direction, are not without their pitfalls. VP of Integrity Guy Rosen refutes the claim that Meta "[has] a financial interest in turning a blind eye to misinformation," citing hubs like the 'COVID-19 Information Center,' the 'US 2020 Voting Information Center,' and the 'reduce, remove, inform' strategy, as evidence of their voluntary, *bona fide* efforts (Meta, 2021). However, Rosen (2021) notes that such efforts come at the expense of user growth and engagement; for example, a 2018 change to the news feed ranking system reduced engaging, short-form content, resulting in a five percent decrease in overall time spent on the platform – in order words, a knack to Facebook's current business model. Rosen also highlights that the platform's enforcement of its policies "will never be perfect," symbolic of the delayed response time of the *removal* component of Facebook's misinformation strategy, citing that misinformation will

[23] Note: "Click farms" are a form of click fraud, denoting a large group of low-paid workers hired to click on paid advertising links to generate ad revenue.

"never be eliminated in its entirety" (Facebook, 2021). While true, going forward, Facebook must emphasize safety over profit to minimize the negative externalities of misinformation all the while bolstering trust and transparency. Doing so necessitates a restructuring of both technocratic interventions and policy aims with the end goal of more *proactive* interventions.[24]

4.1.B Instagram

As a subsidiary of Meta, Instagram also defaults to the 'remove, reduce, inform' approach to combat misinformation, with a few key differences in policy given Instagram's prominence in photo and video content. With Instagram's ease-of-sharing such visual content, navigating the formidable terrain of photo and video manipulation is no small feat, especially with the emergence of advanced deep fake technology – and memes. Nevertheless, Instagram pledges to reduce the spread of false information, defaulting to technology, community feedback, and international, third-party fact-checkers to identify posts and accounts that may contain misinformation (Instagram, 2022).

Via the 'Reducing the Spread of False Information on Instagram' tab in the 'Help Center,' Instagram underscores *explainability* with respect to its intervention policies. Specifically, Instagram outlines its use of the *downranking* mechanism to make misinformation "more difficult to encounter;" in the event third-party fact-checkers identify false information, altered content, or posts with missing context,

[24] Note: The case of Facebook continues in Section 6.1 with a focus on their current technocratic solutions.

such content is filtered from the ‘Explore’ and ‘Hashtag’ pages and “reduced in visibility” on ‘Stories’ and ‘Feed’ pages (Instagram, 2022). To address the challenges associated with visual content, Instagram uses a strategy of image matching technology to detect misinformation-verified content, automatically labeling any identical posts if found elsewhere on the platform. In tandem, Instagram applies false information *label*s*,* which link users to fact-checker ratings and to articles from “credible sources that debunk [the post’s] claim(s),” representative of the platform’s commitment to allowing users the autonomy to decide for themselves what to read, trust, and share (Meta, 2022).[25] Instagram also makes the explicit note that if the content is rated “false or partially false” on Facebook, Instagram will automatically label any duplicate content if it was also posted on Instagram, and vice versa. Lastly, in a similar vein to Facebook, Instagram’s content *removal* mechanism applies to posts or accounts in violation of its community guidelines, i.e., a clear indication of the propensity to induce or incite harm (Instagram, 2022). Taken altogether, Instagram’s intervention mechanisms can be summarized as follows:

	Non-harmful	**Harmful**
Confirmed misinformation	Label	Removal
Disputed accuracy	Label	Removal
Unverified accuracy	No action	No action

[25] See: Appendix Figure 2.

While in theory removing forms of misinformation – especially medical misinformation – represent Kaldor-Hicks improvements, in practice, this intervention has its limitations. Via a content analysis of Instagram during the early stages of the pandemic (April 21-30, 2020), Quin et al. (2021) highlight a potential breakdown in the efficiency of Instagram's removal policy: hashtagging. Specifically, Instagram's removal processes have the potential to miss opportunities for *removal* or *flagging* interventions if hashtags, such as #Hoax, #Plandemic, and #GovernmentLies in the case of this study, are not directly linked. Thus, the cobranding of COVID-19 misinformation with conspiracy theories may allow medical misinformation to spread undetected by removal filters; from there, Instagram's last line of defense is to default to individual users to report or flag content that filters do not catch, including a 'False Information' option when users move to report a post (Instagram, 2022). If posts containing medical misinformation go undetected by intervention mechanisms, this may lead to decreased compliance with public health recommendations or an increased risk of contracting and spreading COVID-19.[26] To achieve a Kaldor-Hicks improvement – where individual users 'win' more than spreaders of misinformation 'lose' – Instagram's intervention mechanisms must also internalize any spillover effects of cobranding.

In addition to cobranding challenges, another potential inhibitor to Kaldor-Hicks efficiency for Instagram is their misinformation policy concerning political

[26] See: Section 5.3 on misinformation's effects on health outcomes.

commentary. Generally speaking, Meta's extensive efforts to combat publicly available misinformation with labeling and removal interventions are commendable, yet they should be expanded to include the direct speech of politicians as well. Meta makes the explicit commitment that the "original content of politicians is not sent to third-party fact-checkers for review," indicating a hands-off approach for political figures (Meta, 2022). While Instagram will not send 'organic content' or campaign advertisements from politicians to their non-partisan 'International Fact-Checking Network,' any otherwise verified misinformation will be reduced in newsfeeds and fitted with *contextual labels* (Instagram, 2019).[27]

In addition, both Facebook and Instagram have had a 'Newsworthiness Exemption' policy since 2016, meaning if someone makes a statement or post which breaks community standards, it will not be removed if the platforms "believe the public interest in seeing it outweighs the risk of harm" (Instagram, 2019).[28] This exception to Meta's fact-checking process has major ramifications for the political process, subjecting the platform to substantial criticism. Meta's decision not to submit direct speech from politicians to fact-checkers is ostensibly grounded in the belief that such discourse is already subject to enough public scrutiny amid the polity and the free press – so much so that further analysis by Meta's fact-checkers is not necessary. As justification, Meta asserts that "in a democracy, people should decide what is credible, not tech companies," while also stressing the important role political ads

[27] Note: The platform works with the Associated Press, factcheck.org, Lead Stories, Check Your Fact, Science Feedback, and PolitiFact.

[28] Note: This does not apply to advertisements, which must always adhere to community guidelines.

play in reaching online communities of voters (Klepper, 2019). This policy essentially gives anyone branded as a politician a platform to lie, mislead, or disinform the public, along with allowing Meta to raise revenue by selling more ads. Thus, any political speech, post, and campaign ad – made by politicians themselves – operate within an entirely different system with respect to Meta's policies on misinformation interventions. Simply put, everyday users of Facebook and Instagram who post misinformation may face the consequences of their interventions, but elected officials are exempt.

Needless to say, this is *not* a minimization of social costs, nor a Kaldor-Hicks improvement; in other words, the negative externalities to society have not been internalized. Retreating from its non-interventionist position in 2020, the platform announced it would "remove posts [from politicians] that incite violence or attempt to suppress voting," in addition to placing contextual labels on posts that violate hate speech policy (Meta, 2020). Nevertheless, the platform's early resistance to implementing welfare-maximizing interventions – alongside the problematic rhetoric of politicians – has conclusively resulted in negative outcomes with respect to democracy, civic engagement, and health.[29]

[29] See: Section 5 for an elaboration on the evidence of misinformation as a negative externality.

4.2 Twitter

"Given that Twitter serves as the de facto public town square, failing to adhere to free speech principles fundamentally undermines democracy. What should be done?"

– *Elon Musk* [30]

As of April 2022, Elon Musk is in the process of acquiring Twitter – after previously amassing a 9.2 percent stake – through a $44 billion buy-out; weeks before, Musk urged Twitter, via the above tweet, to be more transparent, opening up the feed generation algorithms, for one (*The Economist*, 2022). Simply put, Musks's ownership marks a significant advancement in the context of digital platform moderation. Specifically, any future developments for Twitter will be an important case study for how platform leaders can make collective, user-centric decisions as to how individuals want platforms to operate. In other words, Twitter has the potential to serve as a model for other platforms going forward.

As a marketplace of ideas, Musk is right – Twitter most definitely serves as a 'de facto town square,' a forum where the main objective should be to enable a well-informed, vibrant polity rooted in free expression. Musk is also correct in that pursuing the maximization of engagement business model, and subsequently, ad revenue leads to perverse incentives when it comes to intervention mechanisms. With or without Musk, however, Twitter has been taking steps in the right direction to reduce the harms of misinformation, adopting community-driven solutions to boost trust and transparency. Nevertheless, Musks's criticisms prompt the following: Are

[30] Musk, Elon. Twitter Post. Mar. 27, 2022, 1:51 PM.

democratic ideals and free speech at odds with each other? Does a misinformed polity bolster, or threaten democracy? What interventions ensure the greatest number of users benefit, and the lowest number of users are adversely affected? Such questions are the driving force behind Twitter's current and future approaches to misinformation interventions.

As opposed to Meta's platforms, Twitter implements a more regimented approach with its intervention mechanisms; similar to Meta, however, Twitter adopts the following interventions: *removal*, *contextual* and *credibility labeling*, and *downranking* content visibility. Unique to Twitter, the platform enforces a strike system to assign consequences when tweets violate their misinformation policies, conditional on the "severity" and "type" of the violation, along with any previous history of violations (Twitter, 2021). If a user receives one strike, no account-level action is taken, whereas if a user receives two or three strikes, their account will be placed on a 12-hour lock, and a seven-day lock if a user has four strikes; with five or more strikes, Twitter will permanently suspend a user's account, with the possibility for an appeal (Twitter, 2021). With this system, Twitter's intervention mechanisms appear less arbitrary, signaling transparent, yet dependable oversight. A democracy may protect free speech, but that does not mean it does not have rules; according to Twitter, if you violate their rules, you face their consequences.

One of Twitter's rules, in response to an influx of misinformation during the pandemic, explicitly states: "You may not use Twitter's services to share false or misleading information about COVID-19 which may lead to harm," denoted as the potential for increased exposure or adverse effects on health systems (Twitter, 2021). A key component in this December 2021 update is the broadening of Twitter's definition of harmful content, more proactively targeting tweets that directly contradict authoritative health sources like the WHO or the CDC. Also specific to COVID-19 misinformation, Twitter justifies intervention mechanisms if content: "advances a claim of fact," expressed in definitive terms; is demonstrably false or misleading, based on widely available, authoritative sources; and is likely to impact public safety or cause serious harm (Twitter, 2021). Twitter will *remove* specific content concerning COVID-19 misinformation, along the lines of "the pandemic is a hoax" or "5G wireless technology is causing COVID-19," indicative of Twitter's low tolerance policy when it comes to misinformation as a negative externality to society – in the event such content is posted, this will accrue 2 strikes for that user (Twitter, 2021). Given Twitter is largely text-based – as opposed to platforms like Instagram and TikTok – the *removal* mechanism directly implies a Kaldor-Hicks improvement in that the 190+ million Twitter users (*winners*) benefit more than misinformation disseminators (*losers*) are harmed. By listing a myriad of instances where users stand in violation of their policies, Twitter models a high degree of transparency and oversight into when, precisely, the platform will intervene; whether or not this stands in opposition to free speech is for Musk to grapple with.

In addition to *removal*, Twitter will place *contextual* and *credibility labels* onto posts containing verifiably false information. If users are exposed to harm – e.g., high social costs such as adverse public health effects or voter disenfranchisement –

Twitter will prohibit other users' engagement with incendiary tweets to prevent the spread of misinformation. In instances where misinformation does not seek to directly manipulate or disrupt a 'civic process,' but leads to confusion, such content may receive a *contextual* label – if posts containing misinformation receive a label, this will result in one strike to the user's account.[31] Per Twitter's civic integrity policy (2021), violations are grouped into four key areas: misleading information about how to participate; suppression and intimidation; misleading information about outcomes; and false or misleading affiliation. Similar to Meta, Twitter notes that not *all* false information about politics or civic processes constitutes a violation of their civic integrity policy, i.e., Twitter notes a similar approach to 'organic' content made by politicians. Thus, political figures are free to post inflammatory, demonstrably false content – in the absence of other policy violations – that contains debatable viewpoints expressed about elections.

Generally, *contextual and credibility labeling* occurs in the case where authoritative opinion "might change or is changing over time," in situations where local context is necessary, or when the potential for harm is "less direct or imminent," but posts that meet these criteria will *not* receive a strike (Twitter, 2021).[32] For example, labels are placed on tweets that "mischaracterize the nature and science behind mRNA vaccines and how they work" or "misrepresent or misuse official reporting tools/statistics," allowing for public forum-style discussion that guides

[31] Note: Twitter defines civic processes to be events or procedures mandated, organized, and conducted by the governing and/or electoral body; e.g., political elections, censuses, and major referenda/ballot initiatives.
[32] See: Appendix Figure 3.

public conversation toward the accuracy of fact in the absence of imminent harm (Twitter, 2021). These labels, as illustrated in Appendix Figure 3, link users to a Twitter-curated page, or external fact-checking source, containing additional information relating to the claims made in the post. Twitter will also apply 'warnings' – i.e., more direct labels – to a tweet depending on "the propensity for harm and type of misleading information;" e.g., if a post conflicts with CDC guidance (Twitter, 2020).[33] Sanderson et al. (2021), in analyzing how well internalization mechanisms perform, found that 'hard' interventions like *removal* limited the further spread of misinformation on Twitter, but posts that received 'soft' interventions like labels or warnings, spread further than messages that received no intervention at all. Twitter's *removal* mechanism is the primary effective measure for reducing the spread, exposure, and engagement of posts containing misinformation. Acknowledging that misinformation can take on numerous forms, Twitter's intervention mechanisms can be summarized by the following matrix:

	Non-harmful	**Harmful**
Confirmed misinformation	Label	Removal
Disputed accuracy	Label	Warning
Unverified accuracy	No action	No action

[33] See: Appendix Figure 4.

In conjunction, Twitter's interventions explicitly highlight any spillover effects from misinformation, defaulting to content removal when users of the platform are exposed to imminent harm. Given the scope of the pandemic, Twitter has removed thousands of posts containing misleading and potentially harmful content, in addition to intervening in over "1.5 million accounts which were targeting discussions around COVID-19" intending to deceive other users (Twitter, 2020). Given the success of these mechanisms, Twitter's interventions at present should be the operating standard as to how platforms can and should internalize the potential risks of misinformation; given Section 230's ambiguity on *bona fide* actions, Twitter effectively – and efficiently – fills in the gaps.

Albeit well-defined and ostensibly efficient, Twitter's mechanisms are inherently reactionary. Aware of this, Twitter takes the process of internalization one step further. Keith Coleman, Twitter's VP of product, acknowledges Twitter's traditional interventions of *removal* and *labeling* but states that the platform "[doesn't] want to limit efforts to circumstances where something breaks our rules or receives widespread public attention," emphasizing the need for a proactive, community-driven approach to misinformation (Coleman, 2021). In January 2021, Twitter introduced 'Birdwatch,' a U.S.-based pilot program that adopts a user-centric approach to addressing the misinformation problem on Twitter. According to Coleman, Birdwatch "allows people to identify information in Tweets they believe is misleading and write notes that provide informative context," with the end goal of making these transcripts available for the global Twitter audience when there is consensus from a diverse set of contributors (Coleman, 2021).

As of now, user-generated notes are only available on the separate ‘Birdwatch’ website; however, when eventually applied to the platform, the aim is for labeling interventions to be in users’ voices, as opposed to an external or central authority labeling content as true or false. Twitter asserts the development of Birdwatch will be open-sourced and “shaped by the Twitter community” as a whole, bolstering the principles of democracy, as opposed to undermining it –in line with Musk’s general wishes for the platform (Coleman, 2021). All algorithms and data with respect to Birdwatch will be publicly available and downloadable, and the initial ranking system is already made available; the broader and more diverse the group, the better Birdwatch will perform at addressing misinformation (Coleman, 2021). For Musk and Twitter going forward, transparency must be well-defined: it is one thing for people to see the algorithms – or the hard code – themselves, but what people really want to see is the information about *how* they were developed. The computer code gives little insight as to how the models were trained, or what considerations and priorities comprised their development, indicating that making algorithms open source is only a dent in full-scale transparency. Unmistakably, however, the Birdwatch initiative places transparency at the forefront; if this effort proves successful, this would certainly mark a step in the right direction for a restoration of trust, not only at the institutional level but between users as well.

But what does this mean for efficiency? Saltz et al. (2021) found that users – when encountering misinformation interventions overall – found them to be "inappropriate" or "inaccurate," citing negative experiences with *labeling* interventions as a predictor for less support and trust for similar interventions. If Birdwatch proves to be successful in changing attitudes toward specific intervention mechanisms, it has the potential to eliminate any feelings of false positives in misinformation labeling, doing so in a way that underscores trust. With more accurate user-generated context applied to posts containing misinformation, this strategy has the potential to increase the overall 'happiness' of the benefactors, i.e., the users themselves. This is not to imply the complete elimination of negative experiences with misinformation interventions is possible, exploring different ways of addressing the internalization problem is certainly a step in the right direction for maximizing social benefit and minimizing the costs on Twitter.

Looking forward, Twitter holds enormous potential to create the standard for digital platforms when it comes to misinformation interventions. Yoel Roth, Head of Site Integrity, and Nick Pickles, Director of Global Public Strategy and Development, state that "serving the public conversation remains [Twitter's] overarching mission," noting that the platform will continue working to "build tools and offer context" so that users can better navigate credible information (Twitter, 2021). With the addition of Elon Musk – the free-willed arbiter of democracy – to the company, Twitter must remain steadfast in its mission to engage in community-centric interventions, provide a safe space for public discussion, and protect its users from the harms of misinformation.

4.3 TikTok

"Tik Tok was 'just a dancing app'. Then the Ukraine war started."

– Technology Reporter, Kari Paul [34]

Out of all the digital media platforms, TikTok in particular brings the images, videos, and on-the-ground, breaking stories of the world to our fingertips. The content-sharing platform experienced a rapid rise in popularity in 2020, known for turning irrelevant teenagers into millionaires. As of February 2022, TikTok – the top-grossing and number one most downloaded app for the past three years – brought users face-to-face with something entirely new: war. Albeit continuously discounted and deprioritized by those who do not take the time to understand it, now is the time to take TikTok seriously. Simply put, there is no graver externality or threat to social welfare than a senseless war propagated by dictatorial disinformation.

With over one billion users, TikTok, owned by Chinese internet conglomerate ByteDance, is much more than viral dance videos, it's a massive source of information – and misinformation. Put simply, TikTok is a hotbed for misleading content. Unlike platforms like Facebook and Instagram, which mostly present videos and images shared by friends, TikTok's 'For You' function algorithmically generates feed content; i.e., a perfect recipe for going down, and getting trapped within, the rabbit hole. This is problematic in the sense that the more a platform relies on engagement algorithms

[34] See: "TikTok was 'just a dancing app'. Then the Ukraine war started (*The Guardian*).

rather than a chronological newsfeed, the more susceptible it is to mis- and disinformation, given that algorithms favor content that receives more engagement. Regardless, TikTok has been a tool for both Russia and the West to relay information about the war: Ukrainian president, Volodymyr Zelensky, appealed to 'TikTokers' as a group that could help end the war with media literacy efforts; while Russian influencers promulgated the false narrative of "the eight-year genocide" by the Ukrainian people (Mellor, 2022).

According to general manager Vanessa Pappas (2020), TikTok's community guidelines forbid misinformation that has the potential to "cause harm to [the] community or the larger public," specifically misleading content regarding elections or "other civic processes," disinformation campaigns, and medical misinformation (TikTok, 2020). Yet, in 2022, videos of both mis- and disinformation likening Ukraine to a neo-Nazi state were viewed more than 2 million times, feeding disinformation to users whether or not they showed interest in the war (Mellor, 2022). According to an investigation by anti-misinformation organization *NewsGuard*, TikTok is "feeding false and misleading content about the war in Ukraine to users within 40 minutes of signing up to the app," again regardless of whether or not they searched for war-related content (Cadier et al., 2022). TikTok's uncorroborated policies on misinformation do not represent socially efficient optima. So, where does the breakdown in these policies occur?

Inherently, a multitude of features in TikTok's digital infrastructure makes the platform particularly susceptible to hosting and disseminating mis- and disinformation. Above all, anyone can post, or repost, any video without attributing the origin, paving the way for, among other things, the flood of video game content presented as actual footage from the invasion of Ukraine (Evon, 2022). Even if the content is original, TikTok's 'duet' feature, which allows users to react on-screen to the initial post, makes it all the more difficult for users to discern where the initial post originated; with the emphasis on video content, a lot of content screening must be performed in the comment section, which is innately limited. This flawed digital infrastructure makes it difficult for users and journalists alike to discern fact from fiction. Complicating matters further, several TikTok channels are pseudonymous, in that users go by a name other than their own to appeal to a particular audience. Pseudonym-derived usernames create an added challenge when attempting to credit or locate the origin of a post, which is a crucial context for users in determining the originality of videos, especially given the ease of reposting and audio replication.

According to a team of researchers under the Technology and Social Change team at Harvard Kennedy School's Shorenstein Center (2022), who has been monitoring the discourse about Ukraine, there is an absence of "crucial metadata" on the mobile version of the platform. Specifically, the date and time in which a post was uploaded are "not clearly displayed" when users come across videos on their 'For You' feed, finding that users need to take multiple actions – either liking a video or searching for it directly – to verify a timestamp (Nilsen et al., 2022). By contrast, every Twitter post on the mobile or desktop version is marked with a specific date and time, regardless of a user's interaction with the post. Taken together, the existence of these obstacles should be the starting point for how TikTok approaches potential intervention mechanisms. One year before the invasion of Ukraine, TikTok, in a press release by Trust and Safety Manager Gina Hernandez, outlined the platform's intervention mechanisms, summarized as follows: *removal*, *labeling*, and *downranking*. Specifically, TikTok will *remove* misinformation upon confirmation by third-party fact-checkers; TikTok ultimately defaults to the fact-checkers qualifications as to what content ought to be removed (TikTok, 2021).[35] If fact checks are "inconclusive" or unable to be confirmed, especially during rapidly unfolding events when accurate information is difficult to establish, a video "may become ineligible for recommendation into [users'] 'For You' feed," to limit the dissemination of misleading information, indicative of the *downranking* mechanism (TikTok, 2021).

Taking this intervention one step further, TikTok will inform users, via *labeling*, when fact-checkers identify a video with "unsubstantiated content" with the hope of reducing circulation. TikTok users will encounter this intervention via a 'banner,' indicating the post has not yet been verified; if the user attempts to send the video to a friend, they will receive a prompt to pause before deciding to 'cancel' or 'share' the content anyway.[36] TikTok, citing its mission to encourage creativity among its users, argues that when they tested this labeling approach, users decreased the rate at which they shared videos by twenty-four percent, and liked seven percent less unverified content (TikTok, 2021). If only three out of every four users who encounter

[35] Note: TikTok partners with PolitiFact, Lead Stories, and SciVerify to address the accuracy of user content.

[36] See: Appendix Figure 5.

misleading information – and read the label saying it *could* be misinformation – continued to share it, this intervention is by no means efficient; in the case of pollution, if the outcome is such that seventy-five percent of emissions can remain, this would certainly not be 'optimal' for society. As a whole, TikTok's interventions can be summed by the following matrix:

	Non-harmful	**Harmful**
Confirmed misinformation	Label	Removal or Warning
Disputed accuracy	Label	Label
Unverified accuracy	No action	No action

Given this inefficiency, TikTok has been a breeding ground for Russian disinformation. In the wake of the assault on Ukraine, TikTok has scrambled to keep up with the inundation of disinformation about the ongoing conflict – disinformation directly from Russia. Platform spokeswoman Jamie Favazza emphasized that TikTok is responding to the war in Ukraine with "increased safety and security resources to detect emerging threats and remove harmful misinformation," in addition to including digital literacy tips on its 'Discover' page to aid users in evaluating and making decisions about the content they come across (Paul, 2022). Despite these actions, there is much more to be done. Although TikTok suspended its service in Russia on March 6, several Russian-owned and state-run accounts are still visible on the platform; "sputnikvideo," for example, continues to post Kremlin-centric disinformation, albeit with a "Russia state-controlled media" label. These posts can still be viewed or commented on and, at the time of writing, have amassed over 1.3M likes and over sixty-four thousand followers, which directly highlights the ineffectiveness of TikTok's intervention efforts. Internet researcher Abbie Richards (2022) highlights that, as of March 11, a pro-Russia disinformation campaign is using "over 180 TikTok influencers" to promote the invasion of Ukraine, using the caption "Russian Lives Matter" to depict a glorified picture of war and Russian victimization.

Albeit a perfect recipe for the spread of misleading information, 'WarTok' is not all bad news: users inside Ukraine have been using the platform to raise awareness about the conflict and document their first-hand experiences during a time of war, detailing the violence and destruction for the world to see and confront at our fingertips. In addition, users have been taking to the app to fill in the gaps in TikTok's interventions, debunking Russian disinformation campaigns and online rumors.[37]

As such, TikTok does present benefits to society in times of war, but the *risks* of disinformation as an externality – e.g., Russian soldiers using disinformation as justification for burning swastikas into the bodies of fallen Ukrainians, for example – dramatically outweigh any benefit. For TikTok, the path to internalizing the harms of disinformation is simple: *remove* all content that has the propensity to induce harm or to spread an intentionally false narrative. To better approach risk prevention, TikTok should disable the option for users to continue sharing unverified, potentially

[37] See: Users like @valerisssh, @moneykristing, and @xenasolo on TikTok.

disingenuous content, especially when there exists the potential for users to be exposed to harm. Needless to say, TikTok's lack of effective content labeling and moderation, coupled with the algorithmic incentives that push users to content that keeps them on the platform, have made TikTok a fertile ground for harboring and circulating misinformation – more so than platforms like Twitter and Facebook.

Unlike Twitter and Facebook, however, TikTok's front- and back-end infrastructure make it structurally resistant to a trick called 'brigading:' a form of coordinated internet behavior that significantly impacts online safety.[38] No stranger to Facebook, coordinated clickbait farms attempt to get content to go viral, mostly in the name of increased profit through ad revenue. TikTok's 'For You' algorithms create obstacles for brigading, making it difficult for coordinated groups to affect other users' feeds. Also, unlike Twitter or Facebook, TikTok is a 'content platform,' not a social media network; i.e., each 'For You' feed is programmed according to one user's *unique* preferences, relying less on which specific people a user follows. Further, TikTok is a dominated influencer culture, valuing individual creators and microcelebrities who earn a following from a niche group of users, mostly driven by the virality of their content. While the platform is using *labeling* as a primary intervention mechanism, regulating mis- and disinformation on a mass scale appears to become all the more difficult as the power of influencers expands. Unfortunately, this means TikTok influencers play a key role in the proliferation of both factual information and mis- and disinformation, so much so that President Biden held a meeting to brief thirty

[38] See: Social Media Futures: What is Brigading? (Tony Blair Institute for Global Change).

‘top influencers’ on the unfolding crisis in Ukraine (Lorenz, 2022). With the rapid developments of the war, coupled with the desire for primary information, the world is increasingly turning to TikTok; however, in this information frenzy, influencers often cobrand posts with unrelated hashtags and keywords, feeling pressure to stay relevant in the absence of formal accountability (Nilsen et al., 2022). This results in a breakdown of efficiency, as popular content producers have the potential to be plagued by perverse incentives, i.e., doing whatever it takes to gain virality and popularize their accounts.

Thus, when weighing changes to both infrastructure and policy, attention must be given to the demographics who are most susceptible to misinformation on the platform. According to *MIT Technology Review*, it’s Gen Z. Citing her experience as a research assistant at the Stanford Internet Observatory, Jennifer Neda John found that “young people are more likely to believe and pass on misinformation if they feel a sense of common identity” with the initial poster, alluding to how shared experiences and social connections form Gen Z’s collective knowledge (John, 2021). Given that the growing bulk of users on TikTok is between 10-19 years of age, this demographic snapshot is of keen importance (Dean, 2022). With TikTok’s culture of microcelebrities, the platform ostensibly promotes credibility based on identity rather than community, shifting trust to influencers as trusted messengers of information in areas where they have no journalistic credibility. If an influencer can connect with an audience based on shared, personal experiences – childhood bullying, for example – this directly inflates their credibility among younger, niche populations. According to John (2021), this cultivation of “identity-based credibility” signals a future where influencers are “de-facto community leaders” that attract like-minded audiences and affect the distribution of information, which further exacerbates the threat of misinformation. Young people aligned by identity and shared experiences are, therefore, an extremely vulnerable population to misleading narratives that specifically target what brings them together, especially when algorithms are specifically optimized to detect and exploit their cognitive biases.

Given the knowledge of susceptible demographics, structural pitfalls, and the potential for generating harm, TikTok should start by taking its platform seriously. An estimated eight new users join TikTok every second, adding to the total of over one billion users around the world who have an average user session of over ten minutes, i.e., the highest engagement of any digital platform (Kemp, 2022; Verto Analytics, 2019). TikTok has enormous potential to shape collective knowledge; thus, this demands a reevaluation of priorities, especially within feed algorithms. In doing so, ‘For You’ pages should prioritize a diverse mix of content-creators and established news outlets, rather than populating feeds with attention-grabbing content to further increase engagement. In turn, moving feed-generation algorithms away from clickbait will reduce the incentive for the platform, and its subsequent influencers, to generate profit from ad revenue, i.e., the potential to profit off of the spread of mis- and disinformation. Entirely shifting the dynamics of user interactions on TikTok is no small feat; however, the promise of a better-informed citizenry, coupled with the evolving risks of misinformation, should be the driving force behind these changes.[39]

4.4 Google

"Google needs to defund misinformation."

– Professor Noah Giansiracusa [40]

Google – the most popular search engine worldwide with around ninety percent of total market share – connects people to information, shaping perceptions and forming worldviews (Chris, 2022). Yet as a platform, Google is the most dominant company by revenue when it comes to digital advertisements (Mickle, 2021). Thus, the relationship between search algorithms and the monetization of advertisements is of critical importance when evaluating Google's role in the context of the misinformation problem.

In 2016, Google's search algorithms were found to have systematically promoted misinformation on several subjects ranging from climate change to homosexuality (Solon and Levin, 2016). In 2019, the Global Disinformation Index (GDI) estimated that websites characterized by disinformation generated around $250 million in ad revenue, of which Google was responsible for 40 percent (GDI, 2019). In 2020, the GDI found that 1,400 sites spreading COVID-19 misinformation earned a collective $76 million in ad revenue, with Google responsible for over 60 percent of revenues paid (GDI, 2020).[41] In 2021, a *NewsGuard* investigation of over

[39] Note: Important for future study, the influence of the Chinese government on permitting the spread of disinformation, specifically Russian state-controlled TikTok accounts, must be explored in the context of Chinese state intelligence operations.
[40] See: "Google Needs to Defund Misinformation" (November 2021).
[41] See: Appendix Figure 6.

150 websites containing mis- and disinformation about the 2020 election between Election Day in November 2020 and Inauguration Day in January 2021 found that 80 percent of websites containing misleading information received ad payments from Google; to make matters worse, Google also ranked a deceptive *WordPress* blog at the top of search results for the winner of the 2020 election (Skibinski, 2021). Suffice it to say, revenue from advertisements and search engine misinformation are – unfortunately – inextricably linked.

So, what is Google doing, and what *can* Google do, to best internalize the externalities of mis- and disinformation? Concerning digital advertisements, Google, as an intermediary of advertisements, overarchingly aims to enable a "free and open web" where site publishers can monetize content and directly reach online populations, so long as advertisers are within the limits of Google Publisher Policies (Google, 2022). Google's policy on misinformation prohibits misleading content including unreliable or harmful claims, deceptive practices, and manipulated media; in the case of a violation, Google will 'block' ads from appearing on misleading content, and *suspend* or *terminate* the user's account (Google, 2022). For example, in the wake of Russian disinformation campaigns and the ongoing conflict in Ukraine, Google fully suspended all advertisements serving any users located in Russia.

Within its unreliable and harmful claims section, Google explicitly does not allow content that makes "demonstrably false" claims that could: undermine participation or trust in an electoral or democratic process, promote injurious health claims or go against authoritative consensus, or contradict "authoritative scientific consensus" on climate change – an improvement from the blunders of 2016 and 2020 (Google, 2022). Unique to combatting misinformation in the context of digital ads, the *removal* mechanism – *blocking,* in Google terminology – is the most efficient intervention; contextual labels or downranking, for example, do not do much of anything to deter the generation of profit under Google's current business model.

With respect to its search engine, Google outlines three core strategies when it comes to misinformation intervention: making quality count, counteracting malicious actors, and providing additional context to users (Google, 2019). Above all, the fundamental piece to these strategies is the *downranking* intervention mechanism. Under 'making quality count,' Google says it uses 'ranking algorithms' to elevate authoritative, trustworthy information throughout both Google Search and Google News. To best inform these algorithms, Google performs trials with third-party "Search Evaluators" along with "live user tests;" in one year, for example, Google performed over 200,000 such experiments, resulting in more than 2,400 algorithm updates (Google, 2019). When determining the quality of web pages to rank, Google's 'Search Quality Raters Guidelines' explicitly categorize pages that have the potential to impact "future happiness, health, financial stability, or the safety of users," denoting these as, ironically, "Your Money or Your Life" (YMYL) pages (Google, 2019). In other words, YMYL pages are crucial for having an informed citizenry; e.g., medical information, disaster response information, or information about local, state, or national government procedures and policies.

Despite the contention that Google's ranking algorithms are biased toward skewed ideological viewpoints, in practice, Google's *downranking* efforts have been largely successful in reducing misinformation and weighing credible sources more heavily. Google faces a much easier task than platforms like Facebook, for example, in that Facebook must evaluate every single post whereas Google can use a site's track record of harboring misinformation to decide on the proper course of intervention. Taken altogether, Google's search interventions are Kaldor-Hicks efficient given the decrease in the large-scale exposure to mis- and disinformation provides more of a benefit to society than a loss. Similar to Twitter's Birdwatch, Google – via persistent adaptations to its search algorithms driven by user testing – clearly emphasizes a community-based approach to combatting misinformation within the search engine.

In addition to *downranking*, Google, to provide the most appropriate context to users, relies on *labeling* – in the form of 'knowledge panels' – as the main intervention mechanism. In an attempt to side-step the potential for misinformation exposure, knowledge panels appear as information boxes when users perform a Google search for general entities like people, places, or organizations (Google, 2020). The information that appears in the knowledge boxes includes information from a variety of internet sources, fact-check markers to indicate credibility, and feedback buttons for user input. Within individual search results, Google appends *contextual* and *credibility labels* to articles with information by third-party fact-checkers and organizations, in response to searches where resulting claims are demonstrably false or misleading. Putting these mechanisms to the test, Google, after COVID-19 was initially declared a public health emergency by the World Health Organization (WHO), took preventative measures by creating a widespread 'SOS Alert' with evolving information directly from the WHO (Google, 2020). Given the structure and dynamics of Google as a platform and a search engine, providing the greatest number of users with the highest quality information is paramount. As such, Google has steadily increased investment into international preventative initiatives centered around education and prevention.

In 2018, Google launched the Global News Initiative (GNI), to cultivate and foster a "global news community," and increase collaboration between journalists, publishers, and industry leaders (Google, 2022). The initiative seeks to take preventative, *bona fide* action to deter the externalities of misinformation, providing journalists with a myriad of resources and tools while also funding global research and case studies that embolden journalism and information-sharing practices. As of April 2022, the GNI has supported over 7,000 news partners in over 120 countries and territories with over $300 million in total funding (Google, 2022). Within that $300 million, Google has committed $9.5 million specifically to fighting COVID-19 misinformation, supporting fact-checkers debunking potentially false claims, and creating new back-end infrastructure for labeling and downranking interventions (Google, 2022). Not confined to the journalism and information-sharing industries, Google's initiatives rightly devote a great deal of attention to media literacy campaigns, supporting over 15 global media literacy projects as of 2022 (Google, 2022). In addition to the colossal financial commitments to the misinformation internalization problem, Google must also take steps to bolster transparency.

For Google, a principal concern should be to avoid instances of hypocrisy: through ad distributions and auctions, the company funneled money to the same misinformation sites that it simultaneously fights in its public-facing search engine. Looking ahead, there must be a disincentive to this "one step forward, two steps back approach with ad regulation" (Cattich, 2020). Regarding its search engine, Google must be more open and collaborative with key misinformation data points: how many ads have been rejected or removed, how many subject-specific-misinformation search results have been downranked or removed, and qualitative indicators on the types of searches made over precise time intervals. Google is uniquely positioned to collect and collate data on web searches for mis- and disinformation pertaining to democracy, civic engagement, and health outcomes; open-source, explainable data on these markers could better inform researchers, policymakers, and the general public on the origins and flows of misinformation across all digital ecosystems. The Google case, therefore, demands a structural reorganization of the advertisement-based business model when it comes to misinformation interventions. Zeroing in on what party – or parties – directly profit from the advertising value of harmful or misleading content is of the uttermost importance. In doing so, Google's intervention policies will be better informed as to where to devote resources to internalize the misinformation as an externality.

4.5 YouTube

Leading up to the 2020 U.S. presidential election, a now-unlisted YouTube ad titled "Salga a Votar" – posted by Republican nominee Donald J. Trump – that compared Democratic nominee Joe Biden to Venezuelan socialism drew much public attention, most notably in southern Florida. Notwithstanding an Associated Press fact-check, YouTube showed the ad more than 100,000 times in the week leading up to the election (Merrill and McCarthy, 2020). Trump went on to win the state of Florida by roughly 375,000 votes – the largest margin in a presidential election in the state since 1988 – and carried over half of the Cuban American vote. In retrospect, this widespread dissemination of this ad illustrates key gaps in the effectiveness of YouTube's intervention mechanisms. Yet YouTube, a subsidiary of Google, is caught in the crosshairs of anti-politician intervention and advertisement revenue much like Facebook, Instagram, and Twitter. Nevertheless, YouTube adopts a nuanced approach to combatting exposure to misinformation as it stems from video content.

Overarchingly, YouTube makes an explicit commitment to preventing the negative externalities of misinformation, noting that specific types of misleading content with a "serious risk of egregious harm" are not allowed on YouTube (YouTube, 2022). YouTube identifies such egregious harm as any promotion of injurious remedies or treatments, digitally manipulated content like deep fakes, and any intentional interference with a democratic process, for example. Within their policy page, YouTube devotes overt and transparent mentions to civic engagement, COVID-19, and vaccine misinformation policies; here, YouTube offers-up community guidelines videos to illustrate YouTube's stance on these matters and provides specific examples of potential violations.

Within the elections misinformation policy, YouTube outlines the following, non-exhaustive types of impermissible content: voter suppression, candidate eligibility, incitement to interfere with democratic processes, distribution of compromised materials, and election integrity (YouTube, 2022). Under the medical misinformation policy, centered-around COVID-19, YouTube, like several other platforms, does not allow for any misinformation that contradicts authoritative health sources like the CDC or WHO, such as transmission and prevention misinformation. In addition, below the vaccine misinformation policy, YouTube does not permit misleading content regarding vaccine safety, efficacy, and ingredients; on this page, YouTube lists potential examples and includes additional resources to authoritative health sources (YouTube, 2022). Taken together, the mentions of specific forms of misinformation, coupled with various examples, place transparency at the forefront of YouTube's misinformation guidelines, while also underscoring explainability throughout the process.

In response to these specific policy violations, YouTube relies on one central intervention mechanism: *removal*. Specifically, YouTube states that if a user's content violates their misinformation policies, they will "remove the content and send [them] an email," informing the user of the violation and their status within their strikes system (YouTube, 2022). Different from that of Twitter, YouTube's system follows a much stricter progression: the first violation is a non-penalty warning, then any subsequent violation may result in strikes to a user's channel. If a user racks up more than three strikes within ninety days, their channel, and all of their content, will be removed; however, a user's channel will also be removed after a single case of "severe abuse" (YouTube, 2022).

In practice, YouTube relies on both individual users and machine learning algorithms to detect misleading content at scale. Within the 'YouTube Trusted Flagger program,' for example, participants are trained on misinformation policies to provide "robust reporting processes" to policy-facing NGOs, government agencies, and "individuals with high flagging accuracy rates" (YouTube, 2022). Even though the *removal* mechanism is inherently divisive on the topic of free speech (Saltz et al., 2021), YouTube's community-driven approach, unlike other platforms, assumes user oversight throughout the entire review process.

According to the platform (2022), global consumption of potentially harmful misinformation that comes from YouTube's feed recommendations is "below 1 percent" of total consumption; yet this begs the question, is ≤ 1 percent socially optimal?[42] The short answer is no, and YouTube recognizes that by defaulting to third-party evaluators to generate consensuses on the robustness of content. Thus, in addition to *removal*, YouTube, indicates its intervention efforts are guided by the following principles: prioritizing high-quality feed recommendations and recognizing that content monetization is a privilege.

[42] Note: 0.21 percent of views are of violated content that is subsequently removed (YouTube, 2022).

Putting these principles into practice, YouTube integrates the *downranking* and *contextual labeling* mechanisms, depending on whether or not content is verified or 'borderline' misinformation. From there, unlike the other platforms, YouTube relies heavily on machine learning systems to construct models of content review, citing fact-checker consensuses as the main training data.[43] As opposed to *downranking* misleading content, YouTube places more of an emphasis on *elevating* "high-quality information" in feeds, relying again on machine learning systems to "prioritize information from authoritative sources" in both recommendations and search results (YouTube, 2022).

Regarding *contextual labels*, YouTube highlights verified, authoritative-based sources in what they term 'information panels,' which appear as users navigate the platform to provide additional context to give users the autonomy to make their own decisions about what content they watch.[44] In the event of a developing story or crisis, like the war in Ukraine when high-quality video content may not be available, information panels will display links to text-based news articles in YouTube search results, driven by Google's search engine recommendations (YouTube, 2022). Taken together, YouTube's misinformation interventions can be summed as follows:

[43] See: Section 6.3 on the pros and cons of the machine learning approach to interventions.
[44] See: Appendix Figure 7.

	Non-harmful	**Harmful**
Confirmed misinformation	Label	Removal
Disputed accuracy	Downrank	Removal
Unverified accuracy	No action	Label

Undisputedly, YouTube makes clear commitments and takes appropriate action to remove objectionable and potentially harmful content; nevertheless, YouTube is still a significant conduit of mis- and disinformation. In January of 2022, a group of 80 fact-checking organizations signed an open letter alleging that YouTube allows its platform "to be weaponized by unscrupulous actors to manipulate and exploit others," indicative of the influx of criticism directed toward YouTube in the wake of the pandemic (Milmo, 2022). Regarding the "Salga a Votar" ad in 2020, financial motivations, once again, could be at play. In 2020 alone the Trump campaign spent $106 million – $37.2 million in the last month of the campaign – solely on YouTube and Google search ads (Thompson, 2020). Incalculable, however, is the effect running these ads had on disenfranchising or influencing potential Latin American voters in the state of Florida. Given Google's – and subsequently YouTube's – policies on political advertisements, content hailing directly from a candidate's YouTube channels is 'within the public's domain' to question and fact-check; in other words, *not* their responsibility. YouTube spokeswoman, Charlotte Smith states that

YouTube doesn't make any "special exceptions" for politicians, and "Salga a Votar" did not violate their community guidelines at the time (Smith, 2020). So, if YouTube agrees that the potential for harm from medical misinformation is significant enough for removal, what is the hesitation for harm to democracy? Albeit a slippery slope, YouTube must make an unambiguous commitment to *all* negative externalities of mis- and disinformation.

Yet in terms of efficiency, YouTube's principal intervention – *removal* – represents substantial Kaldor-Hicks improvements to public welfare, reducing the potential exposure of mis- and disinformation altogether. But, going forward, YouTube must significantly reduce the response times before removal to avoid increased dissemination. In the case of 'borderline' content where the *removal* mechanism is not applied, YouTube should disable the ability to share such content, along with restricting the embeddability of links to those videos. Given the heavy reliance on machine learning for both content review models and feed recommendations, YouTube must also diversify its models' training data, with more targeted, contemporaneous classifiers in English *and* other world languages. Further, in a similar vein to Google, YouTube should commit to a platform-specific, independent research initiative that looks into how mis- and disinformation campaigns take root and spread on the platform, *especially* in the context of misleading information in non-English content.

4.6 Reddit

Reddit – a platform with over 330 million monthly active users – is markedly different from other digital platforms in that it does *not* have a comprehensive top-down intervention strategy (Feiner, 2019). Rather, Reddit relies on a more localized, people-centric approach to moderating niche communities on the platform. According to Reddit, "every community on [the platform] is defined by its users," noting that some of these users help to regulate the community as moderators – colloquially known as 'mods' (Reddit, 2022). This approach allows the culture, norms, and content to be shaped by the community members themselves, explicitly by the moderators' rules, and implicitly by the up- and downvotes or discussions from individual community members.[45] Above the community-specific rules set by moderators, Reddit, or 'the admins,' set and enforce eight platform-wide rules: no harassment, bullying, or threats of violence; no content manipulation or community disruption; no non-consensual sexually-explicit media; no sexual or suggestive content involving minors; no impersonation; properly label graphic, sexually explicit, or offensive content; no illegal content; and no interfering with the ordinary use of Reddit (Reddit, 2022). In other words, Reddit is like a kindergarten playground for all ages, setting ground rules and delegating the individual post moderation to users themselves.

The emphasis on localized authority and regulatory freedom has allowed for the creation of numerous niche communities and groups to form and flourish on the platform; however, this structure has also laid the foundation that enables mis- and

[45] Note: Depending on the community, upvotes have the potential to imply legitimacy where not warranted.

disinformation to spread with rapid ease. In response to the pandemic, Reddit's admin approach to combat medical misinformation by promoting various resources containing credible information related to COVID-19 if the subreddit is focused on promoting misinformation. With any other subreddit, Reddit admins will preemptively prioritize educating and collaborating with users and mods. If these efforts are unsuccessful, Reddit admins will resort to the following non-exhaustive list of progressive enforcement mechanisms: asking users to "nicely to knock it off," asking "less nicely," temporarily or permanently suspending accounts, removing account privileges, or adding restrictions, placing community-specific restrictions, removing content, and outright banning communities (Reddit, 2022). Within the 'adding restrictions' enforcement, Reddit adopts its own version of the *downranking* mechanism – fittingly regarded as "quarantining" – where once a community is quarantined, it will not appear in search results. If a user attempts to visit a quarantined community, they will be notified that the subreddit may contain misinformation and will have to explicitly opt-in to viewing. In tandem, Reddit also uses banners as a form of *contextual labeling* to similarly promote verified, legitimate content on the platform's homepage and within search results (Reddit, 2022).

Albeit unique to Reddit's mission, how efficient is this semi-hands-off approach to misinformation intervention? Objectively, any hands-off approach that places flagging responsibility, for example, in the hands of individual users does not represent an entirely efficient intervention strategy – if users can still opt-in to viewing misleading information, this does a lot for user autonomy and reducing total exposure, but not for complete internalization. However, given Reddit's infrastructure as an open forum, Reddit can effectively fill in the gaps in moderation and oversight by playing upon its strengths. For example, Reddit is already organizing the "Ask Me Anything" (AMA) series where users can ask questions directly to public officials and medical experts to relay the most up-to-date, accurate information (Reddit, 2022). In addition, given moderators are tasked with regulating their subreddit communities, Reddit must continue to provide its admins and mods the necessary resources and guidance on how and when to conduct specific misinformation interventions. Fortunately, Reddit, in an admin post on safety, made explicit their intention to promptly moderate content that contains any reference to inciting violence or imminent physical harm, such as calls to attack individuals of a specific nationality and suggestions like drinking bleach helps cure COVID-19 (Reddit, 2022).

To specifically address medical misinformation, Reddit has also assembled a set of resources outlining reliable, verified information from health experts for mods reviewing COVID-19 related posts. Standardizing this process, Reddit encourages mods to use the 'AutoModerator' (AutoMod) tool to better identify and remove blatantly objectionable forms of misinformation within their subreddits – AutoMod is a built-in, customizable bot that feeds algorithmic tools to mods to proactively recognize, filter, and remove misleading content (Reddit, 2022). With this tool, mods can select specific parameters – keywords, exact users, website links, etc. – that are not permitted in their subreddit to be programmed into AutoMod. Outside of this tool, Mods can also flag any cases of misinformation or suspicious activity directly to Reddit admins; but on the user level, Reddit states that it will be giving *all* users the ability to report misleading content "shortly" (Reddit, 2022). For all intents and purposes, the Reddit example is a remarkable case study for fact-checking and content review, opting to localize these interventions among a specific group of users instead of taking on moderation responsibilities entirely – or partnering with third-party fact-checkers. Taking all internalization efforts into account, Reddit's mechanisms – or the lack thereof – can be summed by the following matrix:

	Non-harmful	**Harmful**
Confirmed misinformation	Label	Removal/warning
Disputed accuracy	Warning	Warning
Unverified accuracy	No action	No action

Going forward, Reddit's localized approach to content moderation – especially when compared with Twitter's Birdwatch – could set an example for other largely text-based platforms on how to navigate various intervention mechanisms, while doing so in a way that emphasizes user input at every step. To ensure Kaldor-Hicks efficiency in Reddit's digital environment, heightened transparency is key, along with effectively prohibiting any avenue to share verified or potentially harmful content. Reddit should provide periodic updates on all moderation and advertising enforcement efforts as it relates to mis- and disinformation interventions. For example, publishing a routine transparency report, for both mods and users, would highlight when admin intervention has occurred within a specified time frame, thereby increasing information-sharing between admins and mods. Further, such reports should include figures on how many interventions are conducted, how many communities are and have been quarantined, and to what frequency Reddit admins and mods remove mis- and disinformation-related content.[46]

[46] Note: For future study, the platform WhatsApp is certainly worth considering. This case study would shed light on which intervention mechanisms are most appropriate and efficient for *encrypted* platforms.

5 Evidence for Misinformation as an Externality

"This is the cost of disinformation."

– Atul Gawande, in response to a COVID-19 patient's denial[47]

As evidenced by the qualifications of the above intervention mechanisms, the resulting harms of misinformation are equally calculable and consequential. Simply put, the evidence is clear. Mis- and disinformation hinder and undermine the very integrity of fact, with the potential to negatively affect everything from trust in institutions to grounds for war. To build upon the characterization of misinformation as a negative externality, this section provides ample evidence of its adverse effects, highlighting the potential for harm in three key areas: democracy, civic engagement, and health outcomes.

5.1 Democracy

"Political information is to democratic politics what money is to economics; it is the currency of citizenship."

– Michael X. Delli Carpini and Scott Keeter [48]

Extortions to democracy from developments in media are irrespective of the time and state of technology. Throughout the 19th century, inexpensive newsprint and better-quality printing presses allowed partisan newspapers to drastically expand their reach, so much so that Kaplan (2002) argues the effectiveness of the press as a check on power was drastically compromised. Later on, in the 20th century, the emergence

[47] Gawande, Atul. Twitter Post. November 16, 2020, 10:48 AM.

[48] See: *What Americans Know about Politics and Why it Matters* (1996).

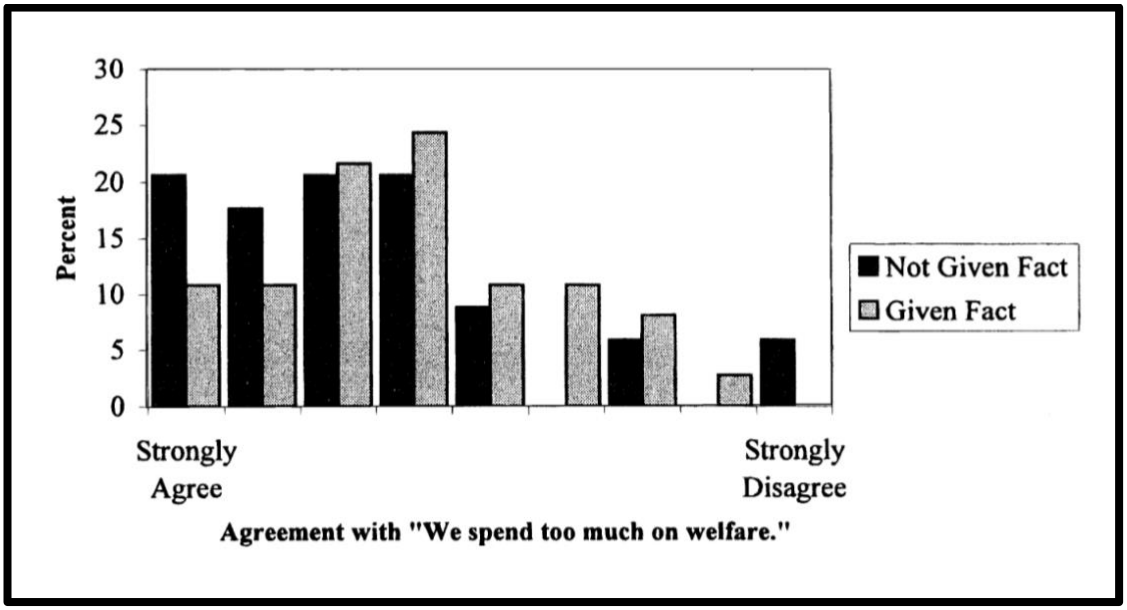
of the radio and television raised concerns that fundamental policy debates would be reduced to sound bites, highlighting charismatic television personalities over more established public servants, all the while creating monopolistic advantages for the largest media conglomerates (Lang and Lang, 2002). In 1968, for example, the Díaz Ordaz administration in México sustained a campaign of disinformation, via radio and television, to intentionally subvert democracy in the wake of the Tlatelolco Massacre (Harris, 2005). At the turn of the 21st century, the proliferation of digital news sounded new alarms, one being the increased and widely available diversity of viewpoints could form large-scale echo chambers; yet in the last 5 years, digital platforms – as harbors for mis- and disinformation – pose different, more direct threats to the very institution of democracy and the ideals that comprise it. Some scholars (Abrahams and Lim, 2020) go as far as to liken misinformation to terrorism in that it prospers in the absence of trust in institutions. Although we have been dealing with the mis- and disinformation problem for centuries, motivations are changing: capitalism drives technological development without internalizing the social costs – or the negative externalities – to democracy.

A functioning democracy depends on a well-informed polity; specifically, concerning *common knowledge* about political actors and the processes they use to earn voters' support (Farrell and Schneier, 2018). Reasonably, the trust held by citizens of a democracy on common knowledge must include that: all political actors act in good faith when vying for office, elections lead to a fair, peaceful transfer of power, and democratic institutions ensure that elected officials exert their power responsibly. When trust in these ideals breakdown – often in the face of mis- and disinformation – the disintegration of democracy follows suit. The incidence of alternative facts and misleading information can and will diminish the trust in common knowledge about democracy, especially when such beliefs become entrenched to the point of robust echo chambers. Evidence of such disruptions in the trust in common knowledge can be found in the 2016 and 2020 U.S. elections, along with the 2016 Brexit campaign where the proliferation of misinformation resulted in large-scale mistrust in voting results.

Given these qualifications for the disintegration of democracy, how can we prove that mis- or disinformation is a driving force? It is one thing to find and isolate data on misinformed citizens, but quite another to demonstrate that misinformation has a causal effect on the political voice of an entire population. Thus, it is crucial to emphasize, instead, the potential for misinformation to skew *aggregate opinion*; more pointedly, the multitude of aggregate opinions, or echo chambers, that dissolve the common knowledge of an entire polity.

Kuklinski et al. (2000), in attempting to show the breakdown of common knowledge assert that not only will people hold *factual* beliefs about public policy, but many also hold *inaccurate* beliefs and do so persistently; in other words, beliefs and preferences are "tightly intertwined" and this combination is what serves as the "barrier to informing the American citizenry." In an attempt to illustrate the causal relationship between beliefs and preferences – i.e., the existence of echo chambers – Kuklinski et al. asked a group of respondents their estimated and preferred levels of public welfare spending, providing only a fraction of respondents with correct information on welfare spending. Respondents who were *immediately* informed of the actual level of welfare spending appeared to ignore or correct their initially mistaken beliefs. Figure 1 compares the aggregate preferences of those who received the correct information with those who did not:

Figure 1: Collective preferences by factual condition



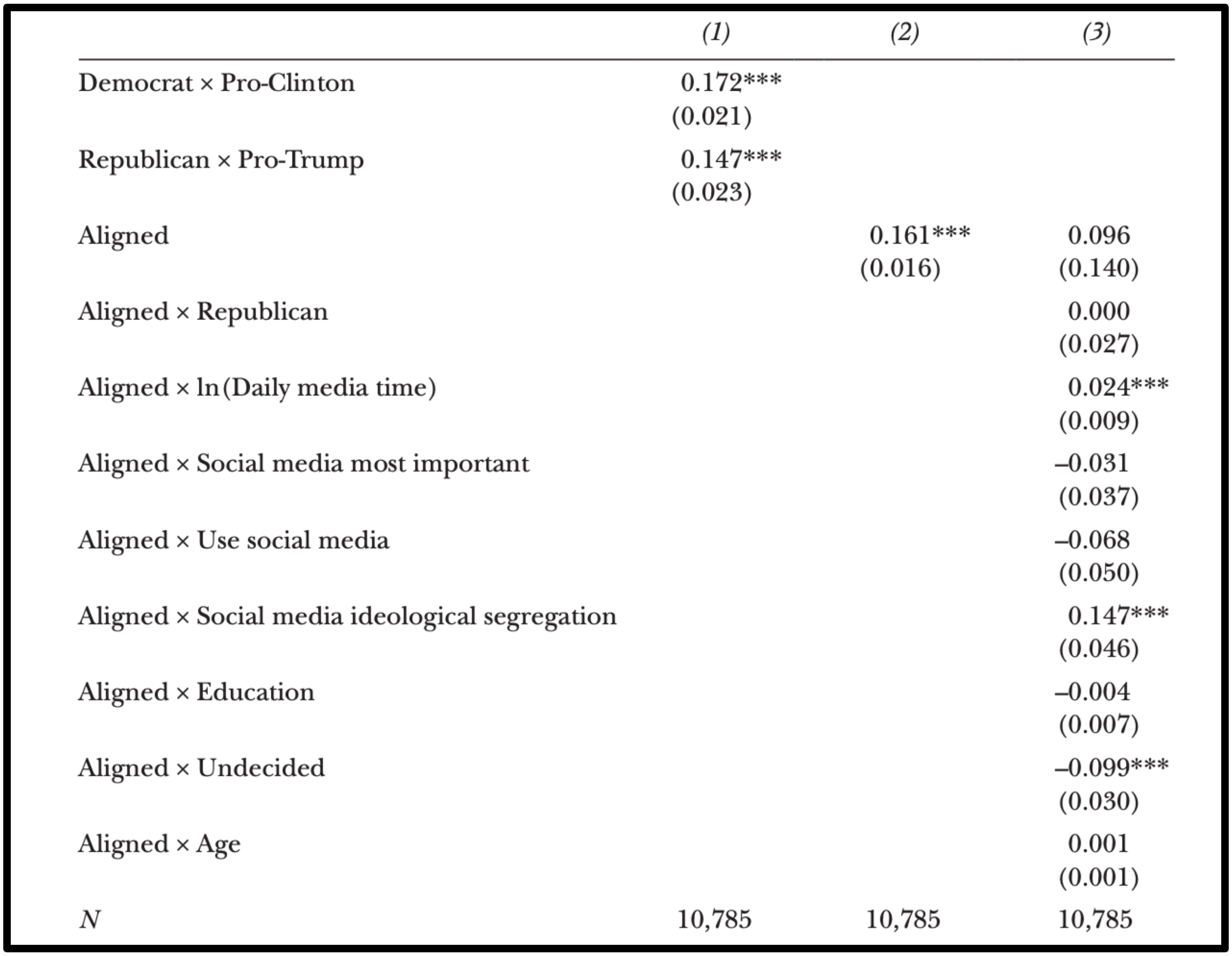
Illustrating the distribution of preferences by respondent group, Figure 1 indicates that respondents who were supplied with factual information generally expressed more support for welfare spending than those who relied on their misconceptions (Kuklinski et al., 2000). As a whole, this study serves to highlight the basis on which individuals make their decisions, whether it be in fact or misinformation, shapes their collective knowledge. These findings confirm the maxim that individuals have strict preferences; i.e., people resist change as opposed to willingly altering their beliefs due to environmental factors. Unless people are, as Kuklinski et al. put it, "hit between the eyes” with the right information, they will continue to judge democratic policies based on their misinformed beliefs. Moreover, even those who are appropriately informed may eventually digress to their original beliefs and preferences, further entrenching them within their echo chamber.

Further exploring the polarization of echo chambers – in the wake of the 2016 U.S. election – Allcott and Gentzkow (2017) investigate any causality between individuals’ ideological alignments and belief in misinformation. To examine any observed effects of echo chambers, Allcott and Gentzkow present the following regression equation:

$$B_{ia} = \beta_{\mathrm{D}} D_i C_a + \beta_{\mathrm{R}} R_i T_a + \gamma_D D_i + \gamma_R R_i + \varepsilon_{ia*} \qquad (5.1)$$

Whereas B_{ia} is a measure of whether an individual believes an article (a), D_i is the democrat indicator, R_i is the republican indicator, and C_a and T_a are indicators for whether headline a is pro-Clinton or pro-Trump, respectively. Per the regression results in Figure 2, Democrats and Republicans are 17.2 and 14.7 percentage points *more* likely to believe ideologically aligned articles than they are to believe nonaligned articles. As evidence of the existence of robust echo chambers, individuals with segregated social networks were found to be significantly more likely to believe ideologically aligned articles. Given the polarizing nature of political misinformation, this may be the case because these entrenched individuals are less likely to receive contrasting information from individuals outside of their echo chambers.

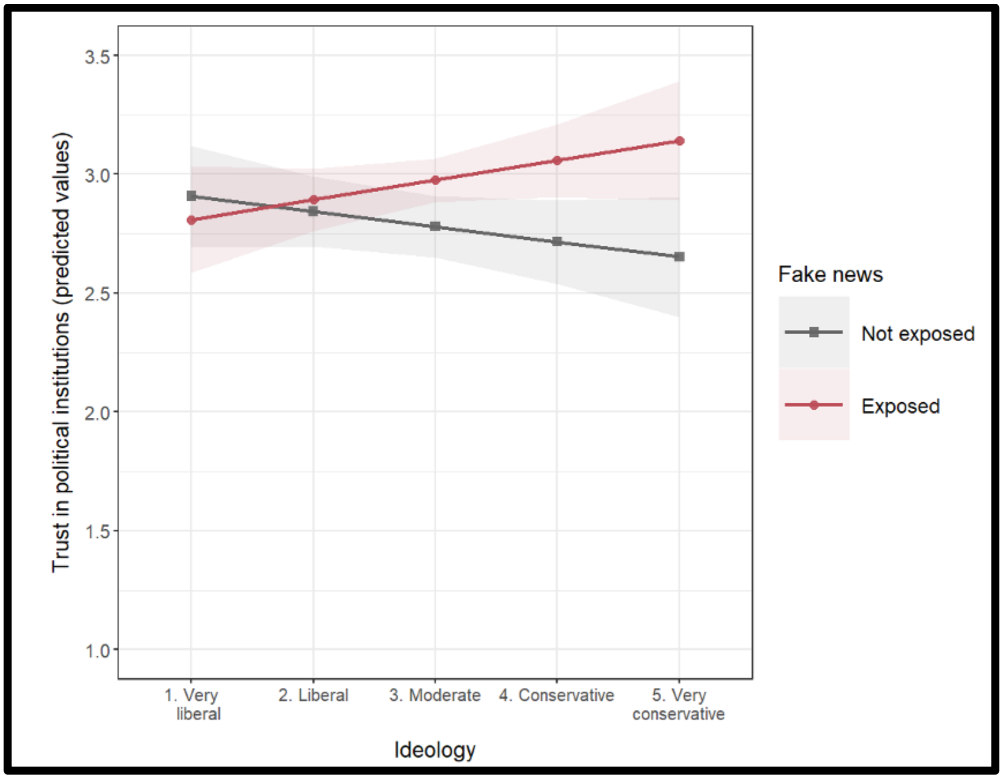
Figure 2: Ideological alignment and belief of news headlines



	(1)	*(2)*	*(3)*
Democrat × Pro-Clinton	0.172*** (0.021)		
Republican × Pro-Trump	0.147*** (0.023)		
Aligned		0.161*** (0.016)	0.096 (0.140)
Aligned × Republican			0.000 (0.027)
Aligned × ln(Daily media time)			0.024*** (0.009)
Aligned × Social media most important			–0.031 (0.037)
Aligned × Use social media			–0.068 (0.050)
Aligned × Social media ideological segregation			0.147*** (0.046)
Aligned × Education			–0.004 (0.007)
Aligned × Undecided			–0.099*** (0.030)
Aligned × Age			0.001 (0.001)
N	10,785	10,785	10,785

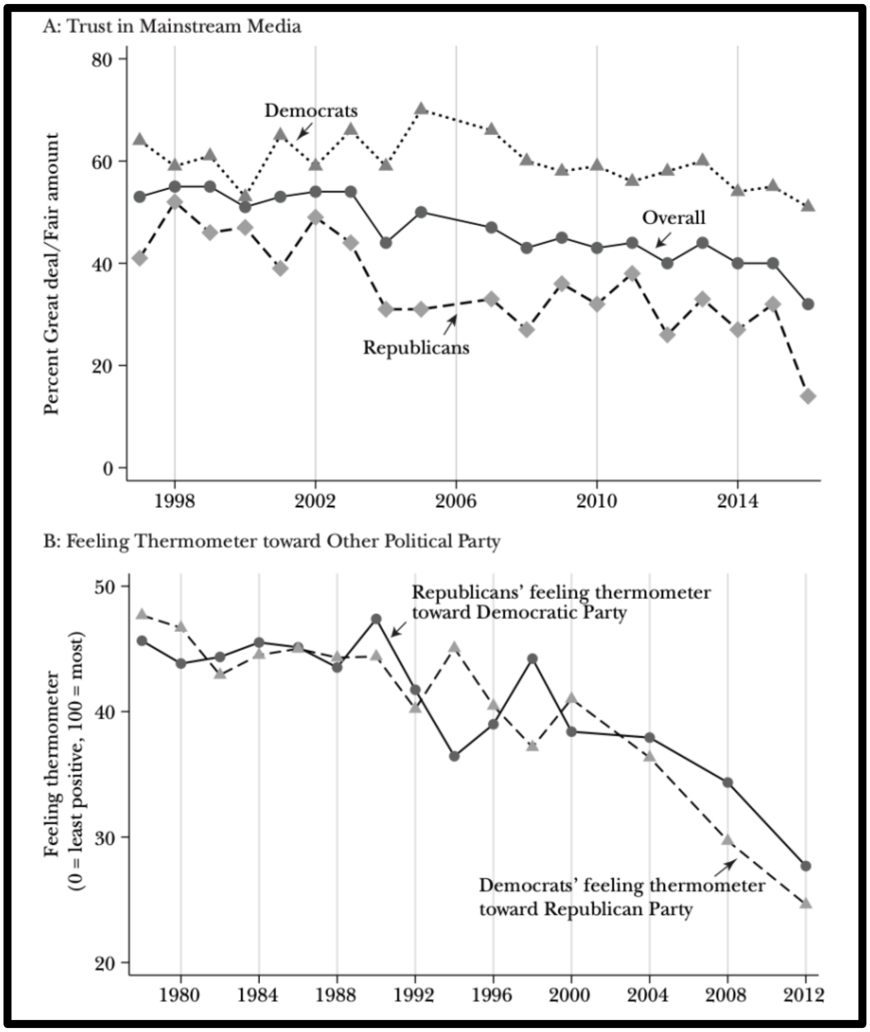
In a similar vein, Ognyanova et al. (2020) examine the possibility that misinformation could erode public trust in democratic institutions, finding that digital misinformation was linked to lower trust in mainstream media across party lines. Interestingly, for moderates and conservatives alike, exposure to misinformation predicted *higher* confidence in political institutions, whereas, with left-leaning individuals, exposure to misinformation had the opposite effect, as portrayed in Figure 3 (Ognyanova et al., 2020). Emphasizing the danger echo chambers pose to democracy on both ends of the spectrum, these findings suggest that while an overall decline in political trust can be harmful, an unsubstantiated increase in public confidence rooted in the belief of misinformation is equally concerning.

Figure 3: Predicted values of trust in political institutions at different levels of ideology and fake news exposure



All things considered, these studies on the declining – and inclining – trust in the institution of democracy carry significant implications. A healthy democracy necessitates two strong, well-informed parties who can, at the bare minimum, be exposed to any opposing viewpoints. With the state of today's stout echo chambers, it appears there is an inverse relationship between trust in mainstream media and political polarization; as overall trust in mainstream media is declining, political polarization is increasing, as revealed in Figure 4 (Allcott and Gentzkow, 2017).

Figure 4: Trust in mainstream media and political feeling thermometer



Given these empirics, the prevalence and proliferation of mis- and disinformation is undoubtedly fueling the disintegration of democracy: creating polarizing echo chambers that play on individuals' confirmation biases that act as a barrier to cross-party information sharing. As former president Barack Obama, in an address at the Stanford Cyber Policy Center, puts it, "One of the biggest reasons for the weakening of democracy is the profound change that's taken place in how we communicate and consume information," yet looking forward, he asks, "Do we allow our democracy to wither, or do we make it better?" (Obama, 2022). Evidently, the choice is up to us.

5.2 Civic engagement

Given the increased polarization of the American polity, civic engagement is of keen importance: a subset of democracy that is adversely affected by the spillover effects of misinformation. An unfortunate, yet prevalent form of mis- and disinformation, *conspiracy theories* – or false narratives – concerning election outcomes surely contribute to the declining faith in the democratic process. Such mis- and disinformation can send the signal to voters that they do not have the potential to affect election outcomes, resulting in all-out disenfranchisement from civic engagement, and subsequently the democratic process. Concurrently, however, mis- and disinformation about the integrity of elections could also *encourage* participation in successive elections, stoking anger and directing believers to continue democratic participation as a means of rectifying their political opponents' perceived misconduct. The latter externality of increased participation suggests that actively spreading misinformation about core democratic functions carries electoral benefits; therefore, the former is more important to emphasize and investigate.

Sustained democratic governance requires that the losing party perceives the winning party as holding power legitimately; if participants cease to believe that democratically-held elections legitimately confer power, democratic self-governance ultimately breaks down (Przeworski, 1999). Leading up to and following the 2020 U.S. general election, then-President Donald Trump, Republican politicians, and several conservative commentators promoted the theory that the election was 'stolen,' perpetuating conspiracies of widespread voter fraud which had no basis in fact (Eggers et al., 2021). Unlike similar rhetoric from past elections, these resurgent claims of voter fraud persisted long after Election Day, morphing into calls to undemocratically overturn the results of the election. The subsequent assault on the Capitol on January 6th, following a rally where Donald Trump continued to pedal disinformation about election integrity, is a painful resulting externality – to civic engagement and for the United States as a "leading" democracy. Bearing these events in mind, an interesting case study on civic engagement – specifically voter turnout and disenfranchisement – is voter participation and changing attitudes surrounding the Georgia senate runoff election that occurred on January 5, 2021.

It is worth mentioning, nevertheless, that in the context of misinformation as it disseminates from digital platforms, any observational analysis that seeks to isolate effects on overall civic engagement is intrinsically limited. Using the 2020 general election as an example, identifying any causal effects of the 2020 election mis- and disinformation on future elections would necessitate simulating large-scale random exposure. In addition, exactly pinpointing any changes in attitudes – or the lack thereof – of individuals' voting choices is also limited. All else equal, the focus should then be on estimating the *likelihood* of voter participation among individuals who either endorse or reject misinformation, as indicated by their behavior on social media – and one study does just that.

Green et al. (2021) examine relationships between public stances on misinformation tied to the 2020 general election and participation in the subsequent Senate runoff elections in Georgia. Their approach tests whether Georgians who "endorsed or rejected conspiratorial content" on Twitter before Georgia's Senate runoff election turned out in that election at different rates than similarly situated Twitter users in Georgia who did not. In doing so, Green et al. leverage a dataset that links Twitter users to voter file records, allowing for comparisons between individuals' behavior on Twitter and their political participation.

Before diving into this analysis, it is beneficial to first touch on the fundamental models of voter turnout to fully understand how widespread misinformation could affect voter turnout. Riker and Ordeshook (2017) specify the calculus of voting as shown below:

$$V = pB - C + B \tag{5.2}$$

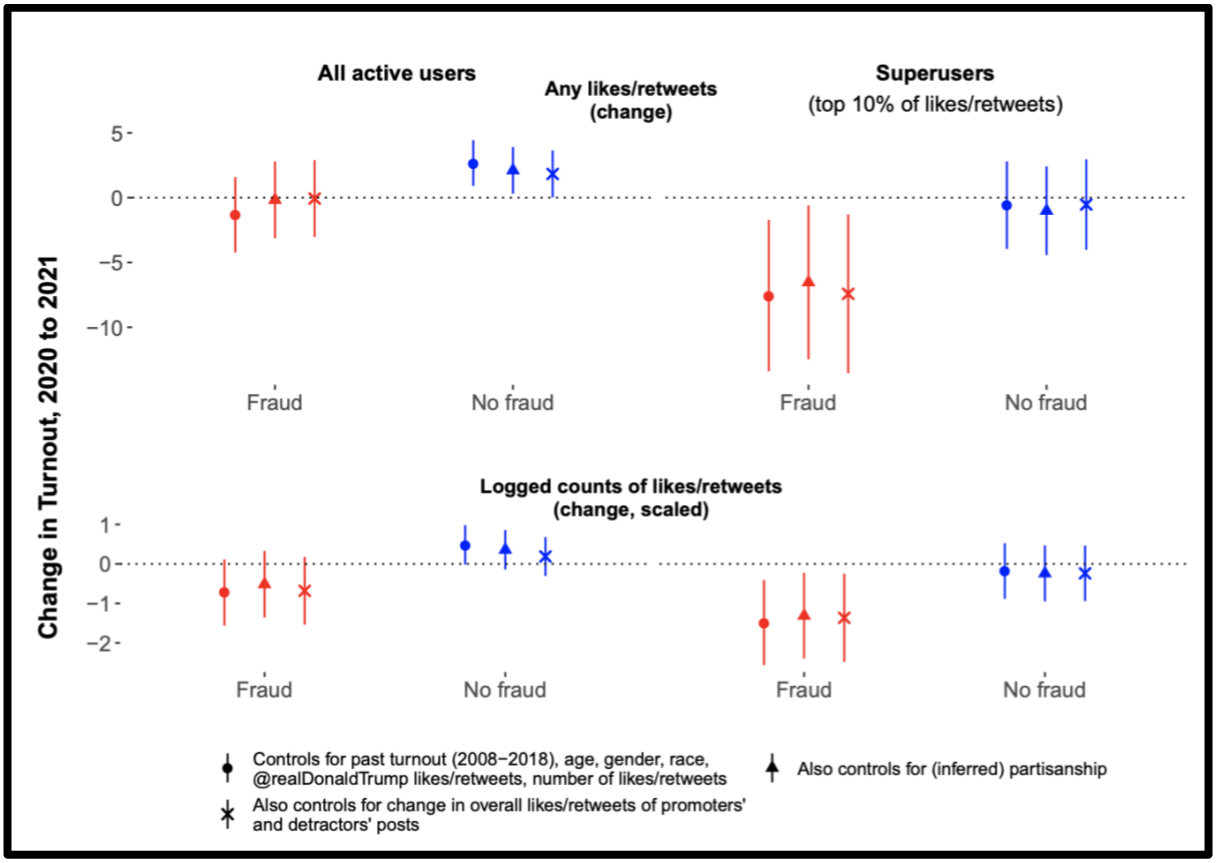
Where V is the likelihood of voting, p is the perceived likelihood that an individual's vote is decisive, B represents the utility associated with an individual's preferred candidate winning, C denotes the costs of voting, and D represents the utility an individual gets simply by voting. According to appraisal theory in social psychology (Lazarus, 1991), individuals' emotions emerge as reactions to particular stimuli and eventually direct people toward particular actions; therefore, claims of widespread voter fraud would increase V among those who believed such claims by inducing anger directed at the opposing party. Additionally, these claims would increase B, the expected utility benefits of winning the election, relative to the costs of losing. Yet, if misinformation concerning widespread voter fraud is conversely framed as an exogenous threat, outside of voters' control, it would then be less likely to produce the sort of anger-induced motivation to turnout. Affecting multiple variables in equation 5.2, misinformation would then weaken the perceived relationship between votes and electoral outcomes; specifically, this would potentially reduce both p and D.[49] This interpretation would be consistent with evidence that the erosion of trust in the electoral process is associated with the decreased turnout among supporters of opposition parties and losing candidates (Domínguez and McCann, 1998; Hernández-Huerta and Cantú, 2021). Thus, if belief in election outcome misinformation is widespread by these qualifications, it would be expected that individuals who endorse such misinformation would vote at lower rates.

Circling back to the Georgia runoff election case, Green et al. (2021) combine behavioral measures of engagement, as expressed on Twitter, and longitudinal voter turnout data at the individual level by linking the accounts of 40,000 Twitter users in Georgia to their respective voter files. Figure 5 displays coefficients from constructed 'change score models' which assess the associations between shifts in likes and retweets of election fraud-related posts before and after the 2020 election with changes in voter turnout in the runoff election minus turnout in the general election (Green et al., 2021). Among 'superusers,' or users among the top 10 percent of likes and retweets of fraud-related misinformation, there was a significant change in voter turnout alongside engagement with such misinformation. This 'subgroup' of

[49] Note: A decrease in p would be the case in a competitive election, whereas a reduction in D would be the case for non-competitive elections.

Twitter users is especially noteworthy as there is more data on their retweets and likes, allowing for a better estimation of their engagement with claims of fraud amid less measurement error.

Figure 5: Change in turnout from 2020 to 2021 and liking/retweeting fraud-related posts



As a whole, Green et al. (2021) conclude that liking or sharing messages opposed to election fraud-related misinformation was associated with a *higher* turnout in the runoff election, whereas among Twitter superusers, those who liked or shared posts promoting misinformation were *less* likely to turn out to vote.[50] Albeit limited in scope, the findings in this study highlight the nuanced potential costs of

[50] Note: Outside the scope of this analysis, other important avenues for study could be a cross-platform behavioral engagement analysis on how the appraisal theory holds in the face of various forms of misinformation.

election-related mis- and disinformation, further characterizing misinformation as a negative externality to civic engagement.

An important behavioral proxy for belief in the legitimacy of elections, voter turnout, and the implications it carries, should undoubtedly be weighed more heavily amongst digital platforms' policies – e.g., using the *removal* mechanism if a post has a high potential to negatively influence or suppress voter turnout. More studies are needed to demonstrate any additional linkages on the overall effects on civic engagement, yet the overarching threat to democracy remains. The perpetuation of mis- and disinformation – especially in the context of microtargeting of false political ads across platforms (Nunziato, 2020) – from the highest levels of government further erode essential democratic norms and threatens democratic legitimacy.

Platforms like Facebook and YouTube, whose lax attitude toward regulation of political speech, further exacerbates the potential for adverse outcomes for democracy like voter manipulation, dissuasion, and suppression. Such lenient intervention policies for political figures essentially give anyone characterized as a politician a platform to lie and mislead the public, without the threat of consequence; all the while allowing platforms to raise revenue by selling more ads on such anti-democratic, yet engaging content. Thus, there must be a mechanistic disincentive to publishing and engaging with such content – across all platforms – for the sake of shielding democratic ideals from attacks on legitimacy.

5.3 Health outcomes

"We're not just fighting an epidemic; we're fighting an infodemic. Fake news spreads faster and more easily than this virus and is just as dangerous."

– Dr. Tedros Ghebreyesus, Director General WHO [51]

Dr. Ghebreyesus is indubitably correct – the COVID-19 pandemic was met with an explosion of inaccurate and misleading information, making it all the more difficult for people to make informed health decisions. During any health crisis, access to reliable information sources and services is critical for the general public to assess preventative healthcare decisions. At the onset of the COVID-19 pandemic, there was a large-scale reliance on digital platforms for emerging epidemiological information, with platform use increasing by up to eighty-seven percent in some parts of the world (Domenico, 2020). This is especially concerning given, that in early 2020, most platforms did *not* have any policies or strategies in place for combatting medical misinformation, allowing for the rapid proliferation of misleading information with respect to COVID-19.

Lee et al. (2020) reported that about sixty-eight percent, i.e., the vast majority, of adults had been exposed to COVID-19-related misinformation through social media platforms or instant messaging platforms in April of 2020. In another study, Naeem et al. (2020) analyzed 1,225 COVID-19 misinformation-related stories from January to April 2020, highlighting that digital platforms, specifically social media sites, accounted for spreading fifty percent of the stories. Concerning preventative

[51] See: Dr. Tedros Ghebreyesus at the Munich Security Conference (February 2020).

vaccination efforts, misleading content about the safety and efficacy of vaccines gained a collective 4.5 billion views on social media in the span of one month (Bryd and Smyser, 2020). Taken altogether, medical misinformation is particularly alarming given the externalities inherent to contagious diseases (Miguel and Kremer, 2004); in other words, the spillover effects of medical misinformation have the potential to impact individuals far beyond those who are directly exposed, affecting disease trajectories and mortality rates in the broader population.

Regarding medical misinformation, as it stems from digital sources, this begs the following question: how does any qualitative divergence in media coverage inform individuals' behavior and beliefs?[52] This question is especially important given the significant externalities involved – e.g., sickness and avoidable mortality – and the consequences of misinformed behavior for the healthcare system as a whole. Again, any analysis that seeks to pinpoint and generalize individuals' behavior or beliefs in aggregation is inherently limited in scope. Thus, to test for any observed effects, we must look at isolated instances of misinformation, or changes in individual media ecosystems. Under such constraints, Bursztyn et al. (2020) present the effects of digressive, or deviating media coverage of COVID-19 by the two most widely viewed cable news programs in the U.S. – *Hannity* and *Tucker Carlson Tonight*.

Bursztyn et al. highlight how differential exposure to information broadcasted through mass media drastically affected individuals' behavior and downstream health outcomes during the height of the pandemic. Airing back-to-back on Fox News,

[52] Note: Divergence from authoritative information sources; e.g., the Centers for Disease Control and Prevention and the World Health Organization.

Hannity and *Tucker Carlson Tonight* had, before January 2020, comparatively *similar* content; however, the two distinctly differed in their coverage during the start of the COVID-19 pandemic. Sean Hannity initially downplayed the severity of the virus, whereas Tucker Carlson – a standout not only at Fox News but also among media colleagues across the ideological spectrum – maintained that COVID-19 posed a serious threat to U.S. citizens as early as February.[53] On February 25, Carlson warned his viewers about COVID-19:

> Currently, the coronavirus appears to kill about two percent of the people who have it. So, let's be generous for a moment and imagine that asymptomatic carriers are not detected and the real death rate is only say half a percent – that would be one-quarter of the current estimates. Even under that scenario, there would still be 27 million deaths from coronavirus globally. In this country, more than a million would die.[54]

By contrast, Hannity covered COVID-19 and its subsequent harms to a much lesser extent than Carlson and other Fox News programs. When Hannity did discuss the virus – especially in February – he downplayed the threat the virus posed; and, in March, Hannity strongly emphasized that Democrats were politicizing the virus, sowing the seeds of polarization. Despite starting to cover mortality statistics in more detail in early March, Hannity still downplayed the threat the virus posed to U.S. citizens; for example, in his show on March 10, Hannity affirmed:

[53] See: "His colleagues at Fox News called coronavirus a 'hoax' and a 'scam.' Why Tucker Carlson saw it differently." *The LA Times* (March 2020).

[54] See: *Tucker Carlson Tonight* (25 February 2020).

> So far in the United States, there have been around 30 deaths, most of which came from one nursing home in the state of Washington. Healthy people, generally, 99 percent recover very fast, even if they contract it. Twenty-six people were shot in Chicago alone over the weekend. You notice there's no widespread hysteria about violence in Chicago.[55]

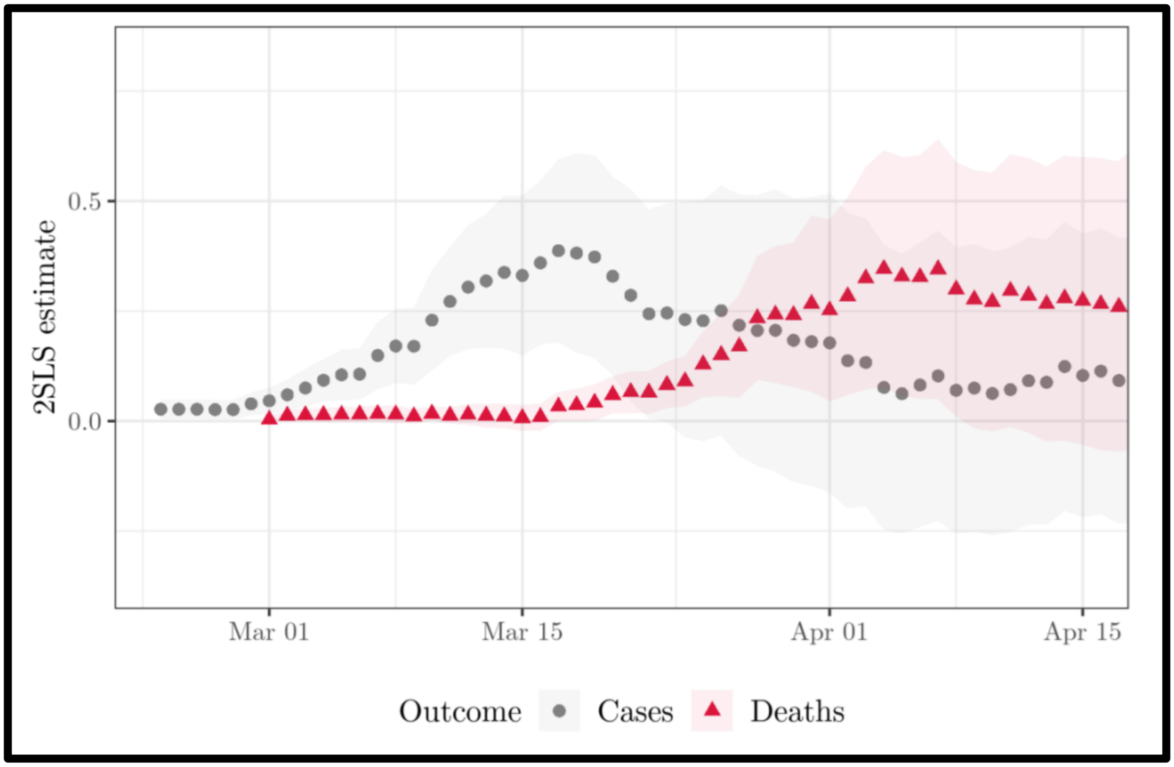
After then-President Trump declared COVID-19 a national emergency in mid-March, Hannity's coverage of the coronavirus converged with that of Carlson and other Fox News shows, highlighting the seriousness of the situation at large and relaying the latest guidance from authoritative sources like the CDC. Nevertheless, these show transcripts illustrate that Carlson explicitly warned his viewers early on about the potential threat that COVID posed, and Hannity did not cover COVID throughout most of February outside of going against public health guidance on potential harms until mid-March.

To corroborate any differences in content, Bursztyn et al. independently codified the shows' transcripts via natural language processing; the team also conducted surveys among the shows' viewers to measure any changes in behavior and belief. Consistent with the differences in content, viewers of *Hannity* were found to have changed their behavior in response to the virus *later* relative to other Fox News viewers, while viewers of *Tucker Carlson Tonight* changed their behavior *earlier* (Bursztyn et al., 2020). Further, by incorporating a 'selection-on-observables' strategy with a robust set of controls and diverse instrumental variable strategies that 'exploit'

[55] See: *Hannity* (10 March 2020).

variation in the timing of TV viewership, greater exposure to *Hannity* relative to *Tucker Carlson Tonight* increased the number of COVID-19 cases and deaths. The day-by-day 2SLS estimates of the effects of the standardized Hannity-Carlson viewership difference on cases and deaths are depicted in Figure 6:

Figure 6: 2SLS estimates of the effect of differential viewership on cases and deaths



As illustrated above, one standard deviation higher viewership of *Hannity* relative to *Tucker Carlson Tonight* was associated with roughly fifteen percent more cases on March 7, thirty-four percent more cases on March 14, and 29 percent more cases on March 21 – statistically significant at $p < 0.001$, 0.001, 0.05, respectively (Bursztyn et al., 2020). Additionally, one standard deviation greater viewership of *Hannity* relative to *Tucker Carlson Tonight* was associated with twenty-four percent *more* deaths on March 28, 35 percent more deaths on April 4, and 30 percent more deaths on April 11 – statistically significant at $p < 0.01, 0.05, 0.10$, respectively (Bursztyn et al., 2020).

Taken together, the rather intuitive results presented by Bursztyn et al. provide quantitative evidence that information exposure is an important mechanism driving the effects in the data. Thankfully, however, the scope of this investigation was confined to the earlier stages of the pandemic before platforms adopted more aggressive intervention mechanisms to combat medical misinformation.

Suffice it to say, these empirics indicate that misinformation on digital platforms and mass media can have significant, and unfortunate, social consequences. Circling back to the opening push-back by Matthew Waxman – "resulting liability for its [misinformation's] harms is very uncertain" – it appears the harms are statistically certain.

6 Remedies

"A society is democratic to the extent that its citizens play a meaningful role in managing public affairs. If their thought is controlled, or their options are narrowly restricted, then evidently they are not playing a meaningful role: only their controllers, and those they serve, are doing so."

– Noam Chomsky [56]

As previously discussed, the decentralized nature of platforms' unilateral policy decisions on misinformation has produced cumulative effects for society at large. Intervention policy at the platform level has effectually partitioned the equilibrium in which wide-subscription platforms, like Twitter, perform a great deal of moderation, whereas localized platforms like Reddit commit to performing significantly less. Thus, when considering potential remedies to internalizing misinformation as a negative externality, it must be acknowledged there is no one-size-fits-all approach – every platform represents a different media ecosystem, and must be treated as such. Above all, however, remedies must be people-centric: *people* are what comprise media ecosystems and *people* have the potential to affect change, at both the ecosystem and individual level.

6.1 Behavioral theory

Labeling the externality problem solely as 'mis-' or 'disinformation' that can be corrected or debunked fails to capture the full scope of the problem. Given the overlap in characteristics between information and mis- and disinformation, as referenced in

[56] See: *Deterring Democracy* (1992).

Appendix Table 1, the externality problem goes far beyond digital platforms in that misinformation has the potential to be *informative*, creating alternative epistemologies and shaping worldviews. Thus, this surface-level framing tacitly implies that, as Lewandowsky et al. (2017) put it, misinformation is a "blemish on the information landscape that can be cleared up with a suitable disinfectant," whereas, in reality, correcting for misinformation necessitates much more than that. But, above all, correcting for misinformation demands a human-centric approach; in doing so, we can leverage behavioral theory to better understand the ways in which people are psychologically predisposed to mis- and disinformation, which will, in turn, better inform any applicable solutions.

At the individual level, then, what cognitive processes are at play with the persistence of misinformation? Albeit a question for behavioral psychologists, individuals' interactions with mis- and disinformation, at the most fundamental level, concern how people assess truth.[57] But what makes people believe certain things over others? Playing on a number of heuristics and cognitive biases, interacting with misinformation and its subsequent moderation interventions often evoke underlying beliefs, stoking anger and thus further entrenching belief in misinformation rather than mitigating it. Lewandowsky et al. (2012) assert that, most often, people are incapable of recognizing that a piece of information is incorrect until they are presented with a correction or retraction. Further, conventional norms

[57] Note: An avenue for future study could be to further investigate the broader "psychology of misinformation," focusing on individuals' behavior in response to various forms of misinformation and intervention strategies.

of everyday conversation favor the acceptance of information as true; namely, conversational information comes with an inferred “guarantee of relevance” (Sperber & Wilson, 1986), and individuals continue under the assumption that speakers are to be honest, relevant, and direct, unless contrarian evidence questions their authority (Grice, 1975; Schwarz, 1994, 1996). Suspension of belief is possible (Hasson, Simmons, and Todorov, 2005; Schul, Mayo, and Burnstein, 2008), yet it demands a high degree of attention, implausibility, or initial distrust of the information received. Given the existence of extremely polarized echo chambers, social-consensus information is particularly daunting to disentangle at the individual level. For example, if Donald Trump and his political allies allege there is widespread election fraud, and they are within your echo chamber, why would you question them? Needless to say, any perceived social consensus can serve to set and sustain individuals’ belief in misinformation.

Concerning the moderation of misinformation, it is necessary to consider how these cognitive underpinnings affect or inhibit the efficacy of intervention mechanisms. Put simply, any ‘correction’ or flag involves a competition between the perceived truth value of mis- and disinformation and the truth value of factual information. Platforms’ interventions should ideally seek to undermine the perceived truth of misinformation, while at the same time enhancing the truth value of accurate information. To do so, however, platforms must attempt to avoid roadblocks such as confirmation biases and cognitive dissonance – the negative experience that follows an encounter with information that contradicts one’s beliefs which can lead individuals to reject any intervention to alleviate the dissonance. Altogether, softening individuals' reliance on their psychological predispositions is no easy task, but doing so emphasizes transparency and trust from individual platforms; hence the community-based approaches to interventions have the most potential in mitigating any alternative epistemologies.

In practice, we can leverage our understanding of the 'psychology of misinformation' to better inform feed generation algorithms. In *Nudge: Improving Decisions About Health, Wealth, and Happiness* (2008), Richard Thaler and Cass Sunstein use a school cafeteria as a thought experiment to demonstrate how young children, much like adults, can be greatly influenced by small changes in their environment. In cafeteria example, the 'director of food services' is presented with a set of organizational choices with respect to how food items should be arranged in a display. Should the food be placed at random? Should the food be arranged to get the kids to pick the same items they would choose on their own? Should the food be arranged such that the students are made best off, all things considered? Or should the goal be to solely maximize profit? Here, the food director is a 'choice architect' – one who has the responsibility for organizing the context in which individuals make decisions (Thaler and Sunstein, 2008). A good architect, in the case of literal architecture, recognizes they cannot build the *perfect* building; rather, they do have the ability to make certain design choices that can have beneficial effects. Choice architects, therefore, can *nudge*.

A nudge, according to Thaler and Sunstein, represents any non-mandated aspect of choice architecture that alters the behavior of individuals in a predictable fashion without blocking or limiting any options; in other words, a form of "libertarian paternalism." Libertarian paternalism – an apparent oxymoron – is an intentionally weak, nonintrusive form of paternalism in that choices are not blocked off or 'significantly burdened' (Thaler and Sunstein, 2008). In the case of the cafeteria, a libertarian paternalistic nudge would be to 'place the fruit at eye level,' meaning that healthier food options are placed in front of less-healthy alternatives, thereby nudging school children to make healthier choices. Contrarily, deterministically banning junk food does *not* count as a nudge. Considering this framework, how can we then recreate a libertarian paternalistic cafeteria within a digital platform?

Regarding feed generation algorithms, digital platforms can use the principles of libertarian paternalism to arrange feeds in a way that does not entirely block off posts, but places the 'fruit' at eye level. In other words, *downranking* verified or potentially misleading content and weighing factual information more heavily; for example, if a user searches "COVID-19" on Twitter, verified information from the CDC would appear first in the feed, nudging users toward more factual sources, but not forbidding posts altogether. Additionally, *contextual labeling* serves to alleviate the concern posed by Lewandowsky et al. (2012) that people generally do not recognize incorrect information until they are presented with a correction. Labels and flags have the potential to nudge – or link – users toward multiple perspectives on specific topics, like COVID-19, allowing users to make the informed choices on what to believe and what to view on platforms.

Given the unilateral nature of intervention policy, individual platforms act as the 'choice architects,' setting and operating within their community guidelines, with every right to determine how their platform is managed. Going forward, platforms can and *should* use the concept of libertarian paternalism as an alternative to censorship; if content does not cross the harm threshold, choices are not blocked off and individual users have the responsibility to make content-viewing choices in their best interest.[58] To avoid censorship concerns from the public, however, platforms *must* be transparent with regard to precisely how feed algorithms are constructed, making such data open-source and open to public inquiry.

6.2 AI and ML: the good, the bad, and the ugly

"*When you're in the business of maximizing engagement, you're not interested in truth. You're not interested in harm, divisiveness, conspiracy. In fact, those are your friends.*"

– *Professor Hany Farid* [59]

Integrating aspects of behavioral theory, platforms by and large, in attempts to curate unbiased, socially optimal feed generations, are turning to advancements in artificial intelligence (AI) and machine learning (ML). Given the rapidly evolving state of technology in the 21st century, we are perhaps getting *too* comfortable defaulting to AI and ML to solve society's greatest anthropocentric problems. AI's

[58] Note: "Censorship" concerns implicitly oversimplify the misinformation problem and often ignore the statutory precedents with respect to limitations on the First Amendment.
[59] See: "How Facebook Got Addicted to Spreading Misinformation," *MIT Technology Review*.

capabilities make it an ostensibly great choice to combat the global water crisis (Chen, 2021), but, in the context of the misinformation problem, we need to move away from technologically deterministic solutions that may inadvertently cause more harm than good. Concisely, platforms must play-on AI and ML's strengths while actively working to mitigate the pitfalls.

Given this upward trend in AI-reliance, we must first follow the money. In 2020, for example, Google provided 'Full Fact' – a nonprofit that provides tools and resources to fact-checkers – $2 million, along with internal Google employees, to help build-out AI tools to better detect misinformation (Google, 2021). Using natural language processing (NLP), or computer programming that allows computers to process and analyze natural language data, these AI-based tools detect claims made by politicians, then grouping them by topic and matching them with similar claims across press, social networks, and radio streams (Google, 2021). With this development in AI, Full Fact increased the amount of claims they could process by over 1000x, identifying and grouping over 100,000 claims per day (Google, 2021). In addition, further developments in NLP will, in turn, allow fact-checkers to spend more time fact-checking misleading information and less time identifying content for review. In essence, ML programs allow identification at scale, making the overall intervention process less time consuming.[60] However, NLP is not without its shortcomings – for NLP to function at scale, it demands a large amount of training data, or thousands upon thousands of examples of misleading information, in order

[60] See: Appendix Figure 8.

to effectively flag. In addition, the vast majority of developments in this space are English-centric; the Full Fact NLP program, as of early 2021, is limited to just four languages (English, French, Portuguese, and Spanish), indicating that, in order for AI to become more internationally applicable, software engineers must extend training data beyond just English.

In practice, NLP tools unequivocally aid the internalization effort with respect to democracy and civic engagement, processing and filtering for familiar language; but what about when language is novel? The COVID-19 pandemic heightened the need to develop new identification tools unique to new and evolving language; in other words, existing misinformation detection datasets could not be applied to pandemic-related misinformation.

Specific to Twitter, Hossain et al. (2020) constructed 'COVIDLies,' a dataset of 6,761 "expert-annotated" tweets to assess the performance of misinformation detection tools on 86 different aspects of COVID-19 misinformation. Publicly available via GitHub, Hossain et al.'s exhaustive dataset of COVID-19 misconceptions is accompanied by tweets that either 'agree,' 'disagree,' or express 'no stance' for each piece of misinformation, contributing to the training data for COVID-19 misinformation detection programs. Using Hossain et al.'s dataset as a case study provides promising evidence for our ability to adapt and evolve alongside emerging events, acknowledging the training data drawbacks and conscientiously working to alleviate them. In addition, the creation of such datasets emphasizes the non-comprehensiveness of digital platforms in that solutions for each platform must be unique to them. Within Facebook, for example, their AI systems successfully filtered for COVID-19-related misinformation and identified any replications, citing the removal of over twelve million pieces of misleading content about the virus and its vaccine (Meta, 2022). Such 'successes' in the AI and ML landscape, however, may be instilling a false sense of confidence and driving overreliance.

So, what happens if AI fails us, and how should we respond? The Facebook case is a prime example. In 2018, the fallout from the Cambridge Analytica scandal – where the personal data of tens of millions of Facebook users was tapped in an attempt to influence voting behavior – put the intersection of AI, ethics, and privacy on full display. This crisis intensified fears that the algorithms that determine feed generations were amplifying misinformation and hate speech, which begs the question: what are Facebook's algorithms coded to do? It's quite simple: to boost engagement. Facebook's algorithms were not created to filter out false or rather inflammatory content; rather, the main objective was to ensure people engage with and share as much content as possible by presenting them with the most stimulating posts and accounts. As Hao (2021) puts it, Facebook's main objective can otherwise be characterized as the "relentless desire for growth," signaling that Zuckerberg would sign-off on anything that supercharged Facebook's expansion. Drawing upon the AI and ML knowhow of Joaquin Quiñonero, a director of AI at the company, Facebook saw AI as the clearcut path to exponential growth.

Applications of ML, however, pose substantial concerns. Unlike traditional algorithms, which are manually coded by software engineers, ML algorithms use training data to, effectually, train itself and ‘learn’ of any correlations in enormous sets of data. A product of this process, the trained algorithm – i.e., the ML model – proceeds to automate future decisions. An algorithm initially trained on engagement data with respect to advertisements, for example, might learn that people in Massachusetts click on ads for winter clothing more often than people in California; the resulting model will then allocate more of those ads to Massachusetts residents. With Facebook’s vast amount of user data points, ML models get all the more detailed and precise. To boost engagement, ads for Canada Goose jackets can be targeted specifically to certain age ranges, zip codes with higher median income, et cetera. While quasi-beneficial for users and advertisers, the reliance on this engagement-centric approach is the root Facebook’s AI problem.

The ML models that seek to maximize engagement, both directly and indirectly, favor misinformation, controversy, and radicalism – i.e., content that enlists visceral, inflammatory or emotional responses from users. This is particularly problematic, given the propensity to inflame existing political tensions on the platform, such as stoking division and violence against the Rohingya Muslim minority in Myanmar which resulted in a full-scale genocide. Even more problematic, Facebook was fully aware: a 2016 internal presentation states that “sixty-four percent of all extremist group joins are due to our recommendation tools,” stemming from the platform’s ‘Groups You Should Join’ and ‘Discover’ algorithms (Horwitz and Seetharaman, 2020). Additionally, in 2017, a task force created by Chris Cox – seasoned Facebook CPO – found a correlation between the maximization of user engagement and political polarization, noting that reducing the likelihood for polarization would result in taking a hit on engagement (Hao, 2021). Since 2017, several Facebook employees have corroborated Cox's findings (Hao, 2021), solidifying the truism that engagement-maximizing models increase polarization. In the five years since Cox's internal findings, little to no substantial changes have been made: Edelson et al. (2021) found that partisan publishers on Facebook received the most engagement in the lead-up to the 2020 U.S. general election and the assault on the Capitol in January 2021.

This lack of action is indicative of the inherent challenges when relying on ML models for misinformation interventions. For example, to catch misinformation before viral spread, content-moderation models would have to be able to accurately identify emerging misinformation. Given ML models build-upon training data, they must be trained on thousands, if not millions of data points of novel content before the model learns what exactly to flag as misinformation; to mitigate this, we need more researchers like Hossain et al. who act fast to construct new datasets like 'COVIDLies' to better inform platforms' models. But these models are still limited, given another potential pitfall is individual users attempting to outsmart ML models by continuously altering wording, or replacing misinformation keywords with pseudonyms.

Regardless of whether or not platforms like Facebook and Twitter use AI or ML-based approaches, people will still continue to spew lies and hate speech. That being said, we should be extremely wary of defaulting to such technologically deterministic interventions, playing on AI's strengths, but also focusing more on the underlying sociopolitical factors that charge individuals' inherent receptivity to misinformation.

6.3 Media literacy

"Even toddlers can understand how not telling the truth, or basing decisions on bad information, can be harmful."

– Peter Adams, SVP at the News Literacy Project [61]

At the time of writing, surveys show that ninety percent of teens aged between 13-17 have used a social media platform; seventy-five percent reporting active use and fifty-one percent reporting daily use (AACAP, 2018). Much like teenagers need health and career planning education, young people also need media literacy education in order to better circumnavigate the externalities digital platforms are proven to induce. Not isolated to younger populations, the need for heightened media literacy programs is only growing for users of all age demographics.

As a result of the influx of disinformation stemming from the war in Ukraine, several platforms have highlighted media literacy resources as a primary mechanism to make users aware of the risks. For example, as of March 2022, TikTok outlined their #BeCyberSmart campaign on Twitter, *nudging* users to consider the reliability of information, as well as the context, before sharing content with others. Additionally, the campaign defaults to 'Media Wise' experts to better inform content

[61] See: "How to talk to kids and teens about misinformation," *MIT Technology Review* (November 2020).

creators with respect to sharing news about Ukraine (TikTok, 2022). Given what we know about Generation Z as an especially vulnerable demographic to spreading misinformation (John, 2021), platforms' media literacy resources must be tailored for users of specific ages. Instagram, in its Help Center, outlines eight general media literacy tips for users to incorporate: consider potential photo manipulation, be skeptical of catchy headlines, investigate the source, watch for unusual formatting, inspect the dates, check the evidence, identify parody accounts, and only share posts you know are credible (Instagram, 2022). As evidenced by TikTok and Instagram, platform-specific media literacy tips are especially important, given the diverse media ecosystems each platform creates. Thus, by outlining explicit tips, platforms are, in turn, shifting accountability and oversight toward individual users; still, platforms can only do so much when it comes to *nudging* users on how to interact within their ecosystems.

Outside of platforms, there exist a number of resources for all age demographics. For example, Kahoot – a popular educational platform for adolescents to young adults – offers up various 'games,' such as "What is 'Misinformation'?' and "What Can I Do About Misinformation?" for young children to get accustomed to the term and how to avoid spreading it before they eventually join platforms themselves (Kahoot, 2022). Karen Douglas, a social psychologist at Kent University, offers up various strategies for parents on how to talk to children about misinformation; e.g., presenting them with a "weak version" of misinformation, then directly debunking it with them, as sugarcoating will not help kids who will undoubtably uncover the truth elsewhere (Basu, 2020).

Adults, too, are no less susceptible to misinformation than children; a Pew Research study in 2020 found that less educated American adults were more inclined to see “some truth” in COVID-19 conspiracy theories, indicating that education at all age levels helps shield people against misinformation. For continuous education at all ages, the Berkman Klein Center for Internet & Society at Harvard University hosts the “Digital Citizenship + Resource Platform (DCPR) – a growing collection of “learning experiences, visualizations, and other educational resources” in areas from civic and political engagement to information quality to privacy and reputation (DCPR, 2022). The resources within the DCPR aim to empower individuals with the knowledge of all parts of the digital world, assisting educators in building media-smart pedagogies with the goal of fostering trustworthy digital spaces for all ages.

For day-to-day use, media literacy efforts are not restricted to education programs and resources. For example, individuals can incorporate accuracy verification tools such as NewsGuard’s that function as desktop browser extensions or mobile applications. As users browse over 7,500 websites, NewsGuard’s accuracy ratings use a color ranking system to highlight the accuracy of news and information pages on search engine results, with green indicating generally trustworthy sources and red indicting generally untrustworthy sources (NewsGuard, 2022). Needless to say, there are countless resources available for internet users of all ages to better navigate constantly evolving, externality-producing digital environments.

Taken together, scholars like Bulger and Davison (2018) and McDougall (2019) vigorously affirm the hypothesis that media literacy is the “center of gravity” for combatting misinformation. Specifically, McDougall (2019)’s ethnographic study explicitly argues that ‘critical’ media literacy – if adopted as a mandatory school subject as part of a ‘dynamic’ education on digital literacy – would better equip young adults and adolescents with the resilience to combat “information disorder” (Wardle and Derekhshan, 2017). Going forward, an acute focus on media literacy pedagogies will serve to emphasize a more *proactive* approach to preventing the externalities of misinformation, as opposed to more *reactive* mechanisms, such as fact-checking and contextual labelling. As Rushkoff (2018) puts it, the former boosts the immune system, while the latter treats the infection.

6.4 Pigouvian taxation

“If you think that moderation [within a firm] is censorship…you’ve got a competition problem.”

– Professor Paul Romer [62]

The tried-and-true economist answer to externality problems – specifically between firms – is a tax, acting as an incentive or disincentive to reach a profit maximizing equilibrium. Paul Romer, the 2018 recipient of the Nobel Prize in economics, proposes a Pigouvian tax to solve the misinformation problem, yet his reasoning is far from traditional, focusing more on platforms’ business models and revenue streams.[63] To

[62] See: “Nobel laureate Paul Romer on how to curb Big Tech’s power,” *UChicago News* (January 2021).

[63] Note: A Pigouvian tax is a tax on any market that generates negative externalities.

Romer, the lack of competition in the tech space is a bigger problem than regulation concerns, shifting the focus to the tech market concentration problem.

Rather than an all-out ban on the engagement-centric business model in which platforms harvest user information to sell targeted ads, Romer proposes a tax that will encourage platforms to shift toward a 'healthier, more traditional' business model (Romer, 2019). Such a tax would be applied to the revenue generated by digital ads; at the federal level, Congress would "add it as a surcharge" to the corporate income tax and, at the state level, it could be adopted as a sales tax on the revenue a platform collects for showing ads to users in that state (Romer, 2019). Aware that technology companies are crafty when it comes to tax avoidance, Romer suggests this is actually a *good* thing, alleging that a tax would spark creativity. To avoid the tax, platforms would have to switch to an ad-free subscription model – much like Elon Musk is proposing for Twitter – where revenue generation is not hinged on compiling data points on users for targeted ads. However, given the annual growth rate of global advertising is hovering at around twenty percent (Axios, 2021), platforms would most likely still pursue the targeted ad model in order to yield higher profit. Nevertheless, Romer suggests that in order to solve the competition problem, such a tax should be *progressive*, in that the tax rate will increase the larger a company becomes, further disincentivizing concentration in the tech market.

Given that Romer won a Nobel Prize for his work on technological progress and economic growth, when someone with his credentials calls Big Tech a "threat to our social and political way of life," his proposals are worth serious consideration.

However, the imposition of a Pigouvian tax, while fitting for traditional negative externality-producing markets, will most likely do very little to internalize the externalities of misinformation. A progressive tax may serve to realign priorities away from the engagement-based, advertisement model, but this would result in indirect and direct costs – paying for a subscription, for example – to the individual users themselves. Theoretically speaking, the problem with using a Pigouvian tax to combat misinformation is that we would need to know the 'optimal level' of the externality (misinformation) in order to impose any tax. Further, another fundamental problem with using a Pigouvian tax is that there is a missing market: the market for misinformation.[64] While it is theoretically possible to construct a market – as the next section attempts to do – such a model would be riddled with assumptions, given the spillover effects of misinformation are often difficult to measure empirically. While the underlying motivations in Romer's proposal are presumably valid, levying such a tax is improbable and does not directly target the misinformation problem.

Aware of the inherent pitfalls of Pigouvian taxation, Romer, at an antitrust conference held at UChicago Booth, offered-up other potential mitigation strategies: public records of all platform advertisements, more transparent information about how many and what types of people ads are shown to, and a more flexible merger-review process (Kasperkevic, 2021). If Romer's proposal tells us anything, it is that combatting the misinformation problem requires multi-dimensional solutions. There

[64] Note: An avenue for future study would be to postulate a thought experiment in which misinformation levels across platforms functioned similar to a carbon cap-and-trade system.

is no one-size-fits-all approach, but rather a combination of approaches addressing each player in the media ecosystem.

6.5 A mechanism design approach

In an attempt to simulate the incentivization of interventions, Dave et al. (2020) construct a resource allocation mechanism between two players: digital platforms, motivated by capitalistic values, and a strategic government. Within this game, platforms are assumed to be motivated by maximizing their advertisement revenue, whereas the government, via a 'social planner,' seeks to make an investment to incentivize platforms to filter for misinformation in order to reach an optimal level of intervention. As qualifications for this game, Dave et al. assert that such a mechanism must incentivize platforms to *voluntarily* participate in the game, and induce a selection of interventions that maximize social welfare – defined as the sum of utilities for all platform users. Albeit an assumption, a fundamental pitfall of this induced game is the notion that individual platforms will *voluntarily* participate; thus, a more representative game between platforms and the government should assume *involuntary* participation, given the liability obligations of Section 230 reform.

Nevertheless, Dave et al. construct the following optimization problems, platform *i*'s and the government's, respectfully:

$$\max_{m_i \in M_i} \ v_i\,(a_k\,(m) : k \in C_i) - T_i\,(m) \tag{6.1}$$

$$\text{subject to: } 0 \leq a_i\,(m) \leq 1 \tag{6.2}$$

$$n_i\,(m) - n_i \cdot h_i\,(a_i\,(m)) \leq 0 \tag{6.3}$$

Whereas 6.1 represents the utility u_i (m_i, m_{-i}) of platform *i, v_i* is the valuation function of player *i, a_k* is the filter proposed by a platform for platform $k \in C_i$, and T_i (m) represents the tax allocated to player *i*; 6.2 ensures the feasibility of the allocated filter a_i (m); and 6.3 ensures that the allocated 'minimum average trust' is achieved (Dave et al., 2020).

$$\max_{m_0 \in M_0} \; v_0\,(a_0\,(m)) - T_0\,(m) \tag{6.4}$$

$$\text{subject to: } 0 \leq a_0\,(m) \leq 1 \tag{6.5}$$

$$\pi_0 \cdot a_0\,(m) - b_0 \;\leq 0 \tag{6.6}$$

Whereas 6.4 represents the utility u_0 (m_0, m_{-0}) of the government*,* T_0 is the tax allocated to player *i;* 6.5 ensures the government's lower bound a_0 is feasible; and 6.6 is the budgetary constraint on total investment (Dave et al., 2020). Noting quasi-concave utilities, Dave et al. suggest this mechanism arrangement implements a Pareto efficient solution; however, a potential pitfall to this baseline for efficiency is that, with any degree of assumed perfect competition in a market, *any* equilibrium will then tend to be Pareto optimal. These conditions would assume that a Pareto improvement is *not* possible via government intervention; effectually, any intervention that would make platform *i* better off would make another platform worse off. Additionally, equation 6.1 clearly emphasizes the utility of individual platforms, $\sum i$, yet a more optimal outcome variable to emphasize is engagement, i.e., minimizing exposure of misinformation rather than maximizing the utility of platforms' interventions. In doing so, the baseline for efficiency would be more representative of Kaldor-Hicks improvements as it focuses on the utility of users, rather than the utility of platforms.

Dave et al. use the above assumptions to construct the following thought experiment: the interactions between three platforms – Facebook, Twitter, and Reddit – and the government. Here, user engagement is assumed to be the primary driver of advertisement revenue, and each platform seeks to optimize their feed recommendation algorithms in order to maximize users' engagement. In practice, each platform is assumed to have the ability to intervene, via flagging mechanisms; however, filtering is presumably expensive due to the cost of identification and the resulting decrease in user engagement. Thus, the government would allocate a budget for the misinformation problem and appoint an "independent agency" to "design monetary incentives" for platforms, inducing voluntary participation that maximizes social welfare (Dave et al., 2020).

Through the 'participation' and 'bargaining' steps of the mechanism, the platforms and the government would then reach a consensus on acceptable levels of misinformation filtering; i.e., a 'generalized' Nash equilibrium (Dave et al., 2020). These conditions would represent an equilibrium as long as the government commits to addressing the problem of misinformation. Thus, the allocation mechanism would ensure that platforms will ultimately agree to implement filters; from there, the government's 'investment' is subsequently distributed to Facebook, Twitter, and Reddit as a subsidy after they implement the filters.

Taken as a whole, this proposed game would be entirely nullified if participation was assumed to be involuntary; thus, refinements to the modeling framework are crucial, concurrent with applicable statutory reform as outlined in the following section. Albeit technically sufficient, this model neglects to emphasize the true bearers of the costs of misinformation: individual users. A reinterpretation of such a mechanism should seek to maximize the utility, or social welfare, of individual users via a move toward Kaldor-Hicks efficiency as outlined in Section 3.2.

6.6 Statutory modifications to Section 230

"Policy is powerful. Governments have the power to do enormous good but also important damage."

– Abhijit V. Banerjee and Esther Duflo [65]

When looking at a set of producers and consumers, market mechanisms can theoretically determine supply and demand, with the end goal of achieving Pareto efficient allocations. But, in the presence of externalities, the market will not *necessarily* result in Pareto efficient provisions of resources. However, there are alternative methods of recourse such as the legal system or government intervention that can mimic market mechanisms to some degree and thereby achieve efficient outcomes for both parties. In the case of misinformation, the mechanism is Section 230. Here, Banerjee and Duflo are right again: governments have the power to do "enormous" good, but the absence of Section 230 reform will inevitably prolong and sustain the "damage" of misinformation.

[65] See: *Good Economics for Hard Times* (2019).

Passing judgment on the regulation of free speech entails a balancing act: upholding First Amendment protections while promoting the health and safety of individuals the constitution aims to protect. In some cases, the right thing to do from a safety or security standpoint is not the most ideal for privacy or free expression; thus, there will *always* be a tradeoff between disappointing people and promoting the common good (Cattich, 2020).[66] Deciding upon such tradeoffs is anything but straightforward and the decentralization of platform moderation, in turn, raises much ambiguity surrounding the enforcement of Section 230. Given the relative ease of switching between media ecosystems, amending the clauses of Section 230 is of utmost importance – working *with* platforms, rather than against them.

Given the ambiguity of Section 230 subsection (c) – as outlined in section 2.2 – a modification to the statute is long overdue yet calls to modify Section 230 are nothing new. Contemporaneous work in this field calls for a reinterpretation and a reassessment of Section 230 altogether. Lotty (2020) argues the current interpretations of the scope of Section 230 immunity wrongfully deny individuals who have been sexually harassed or assaulted an opportunity to hold platforms accountable for causing or exacerbating their harms. Thus, Lotty prescribes a revision to subsection (c) that "narrows the scope of immunity" and clarifies ambiguities. Specifically, subsection (c) demands a narrower interpretation of what

[66] Note: ethical considerations arise when the onus falls upon companies, or any governing body, to determine right from wrong. Defaulting to individual platforms for moderation effectually limits the role and presence of government in our daily lives.

constitutes a "publisher," and a regression to industry standards to characterize "Good Samaritan" behavior.

In a similar vein, Sloss (2020) responds to the Myanmar military carrying out targeted digital attacks via Facebook against Rohingya Muslim communities in Rakhine State in 2017. Sloss argues for a liability *exception* in subsection (c) to permit civil suits against digital platforms on account of their alleged complicity in genocide, war crimes, or crimes against humanity. Given a civil liability exception in this case, Rohingya plaintiffs could then bring a state tort law claim against Facebook, alleging Facebook's negligence in allowing its platform to become a catalyst for calls of mass violence against an entire population. As the clauses in Section 230 stand today, any such case would be immediately nullified under the current federal preemption defense to state tort law claims.

Irrespective of the study or the context, the written law – or the lack thereof – in Section 230 certainly demands a modification to liability protection; specifically, a modification that provides a positive incentive for digital platforms to internalize the risk prevention of misinformation. Most recently, Congress passed a bill in 2019 that added a new exception to Section 230 that positions platforms to be liable for any third-party content that facilitates sex trafficking on their services (Goldman, 2018). Building off of this addition, subsection (c) must make clear what constitutes "any voluntary action in good faith" to provide collective guidance as to what intervention mechanism platforms should adopt, and under what specific circumstances.

To touch on these points, Congress should draft a third paragraph in subsection (c) titled, “Exceptions to Civil Liability Protection” which explicitly permits civil liability in the event digital platforms are found to be complicit in the circulation of misinformation. Under subsection (c), such an exception to liability would arise in the case that a company fails to prevent transmission of inaccurate information if that information: a) would be understood by “average” readers as incitement or inducement to mislead or suppress people of factual information and the written law; and b) there is a significant risk that any recipient of the inaccurate information is exposed to harm (Cattich, 2020). Specific mention must be given to potential harms in the areas of democracy, civic engagement, and health outcomes, citing the evidence from Section 5.

Within this subsection, there must also be a well-defined time frame during which platforms are obligated to remove posted content in violation of the statute to incentivize faster reaction times; e.g., twenty-four hours since the post’s inception. Make no mistake: the inclusion of “Exceptions to Civil Liability Protection” will continue to preserve immunity from civil liability for digital platforms that make a reasonable, good faith effort to comply with content removal policies, i.e., in the event interventions ultimately exceed the agreed-upon time window. To do so, however, such *bona fide* efforts must be well-defined in subsection (c), paragraph (2)(A).

Ideally – and historically – it is the work of journalists to hold those in power accountable and to provide context to readers; however, given the massive scale and reach of platforms like Twitter, Google, and Facebook, the clauses in Section 230 *must* incentivize action. Thus, to define what proper good faith action is in practice, the legislation should recommend various aspects of platforms' current approaches to best outline what voluntary steps could look like for all platforms. At the most basic level, interventions must: require platforms to take action to limit the potential for viral spread of misinformation through algorithmic adjustments, flag confirmed inaccurate, but not imminently harmful, information labeled with third-party fact checks, and allow content to be seen such that people can debate, support, refute, or dissect it, and to do so publicly.[67]

6.7 Why 'revoking' Section 230 is *not* a remedy

"Revise, revoke, repeal – whatever 'R' word you choose – the debate over the future of Section 230 boils down to one, basic concept: the law exists to protect people."

– Ryan Cattich [68]

As recently as December 2020, former president Trump threatened to veto an annual defense bill unless Congress "revoked" Section 230 of the CDA in its entirety. One year prior, President Joe Biden, in a December 2019 interview with *The New York Times,* pledged that "Section 230 should be revoked, immediately be revoked, number one," highlighting the strange, yet misaligned contention between the left and the right with Section 230.[69] Specifically, the left sees Section 230 reform as key to reducing the harm of online misinformation, whereas the right sees reform as a way

[67] Note: a potential avenue for future study would be to evaluate the various court interpretations of the term "publisher," as it either relates to or differs from the term "platform" in both legal provisions and court decisions.

[68] See: "The Digital Pandemic" (2020).

[69] See: "Joe Biden: Former Vice President of the United States," *The New York Times* (January 2020).

to prevent censorship. In the wake of Elon Musk's Twitter buy-out in April 2022, calls to "revoke" Section 230 have resurfaced. So, let me be clear: all-out revoking Section 230 is *not* the solution.

Above all else, Section 230 serves one main purpose: shielding digital platforms and social media companies from liability for what individuals do and say on their platforms. If Section 230 were to be revoked – as Trump and Biden, and both the left and the right have called for – it would create a catastrophic editorial dilemma. Platforms of all sizes would have to effectually over-censor and remove any content that might *remotely* run a liability risk, causing a ripple effect in the timeliness of the internet and imposing substantial costs, which users themselves would ultimately bear. Essentially, platform-specific legal teams would have to approve of any content before publication on digital platforms, involuntarily making the largest companies, who can afford large-scale legal support, all the more powerful, thereby exacerbating the Big Tech concentration problem Paul Romer warns about.

Section 230, albeit not perfect, is essential to allowing platforms to exist within their discretion to perform intervention mechanisms as they see fit. Protecting everything from Facebook to 8chan, the central goal of Section 230 is to provide platforms with the certainty that they can adopt particular, good faith moderation practices that users deem necessary – without the constant risk of liability. In the absence of this civil liability protection, the entire internet ecosystem as we know it would look vastly different, with some platforms shutting down and others stopping moderation altogether (Cattich, 2020).

So, how can we, as lawmakers, policy drafters – and dare I say, economists – orient the clauses and incentives of Section 230 such that the law exists to protect *all* people? Integral to the main objective of a Section 230 modification is ensuring more reliable and truthful content appears, spreads, and remains on online platforms. It would be wrong to assume this is solely a debate over the future of free speech; thus, pleas to revoke Section 230 in totality are both unsubstantiated and grounded in misaligned priorities (Cattich, 2020).

Platforms across the board are continually innovating with new forms of intervention mechanisms to contest the spread of misinformation, yet they are plagued by the subjectivity of their own policy decisions. Thus, there exists a deep desire for heightened collaboration and information-sharing between the tech and media industries and the U.S. government. T]=he path of least resistance in determining how platforms should be governed is to resolve these tensions and ambiguities together as a society, in a reasonable and equitable fashion.

7 Looking Ahead: Where do we go From Here?

"Social media is a tool. At the end of the day, tools don't control us, we control them. It's up to each of us to decide what we value and then use the tools we've been given to advance those values."

– Barack Obama [70]

On April 21, 2022, former President Barack Obama spoke at the Stanford University Cyber Policy Center, making incredibly pointed remarks about the current state and future of the technology landscape. In his speech, Obama touched on – almost – every point made in this analysis, emphasizing, above all, that human-centric problems do, in fact, have human-centric solutions. At one point in the speech, Obama noted that the companies behind digital platforms "need to have to some other north star other than just making money" and expedited growth (Obama, 2022); taking that one step further, there must also be a parallel 'north star' reorientation at the individual level. Irrespective of echo chambers, we must take a collective step back and consider what *we* truly value as a society: what we desire from the platforms we subscribe to and what we want for the future generations of platform users. Doing so – as Banerjee and Duflo assert in *Good Economics for Hard Times* – demands a profound restructuring of socioeconomic priorities.

To better inform our policies, our laws, and our governance, we must undoubtedly place human dignity at the forefront. Economic efficiency, largely rooted in profit maximization is not as black-and-white in practice, and one economist, or

[70] See: "Challenges to Democracy in the Digital Information Realm," (Obama, 2022).

governing body, does not have the power to define efficiency for all people. Hence, my definition of efficiency – the quasi-utilitarian approach to the Kaldor-Hicks definition – is most definitely not the be-all, end-all. Instead, our notion of efficiency will continue to evolve much like we do, but the focal point should never waver. Minimizing social costs, maximizing social benefit, ensuring the winners win more than the losers lose, or however you want to classify it, serve to emphasize the role we, as individuals, play in the grand scheme of policymaking. The very institutions we subvert to all share a common thread – *people*. People comprise institutions, userbases, data points, and communities, both on- and offline. Acknowledging this truism, how can we reflect this in the context of the misinformation problem?

Another way to think about misinformation, according to Dartmouth Professor Brendan Nyhan, is that it's not about us. Namely, platforms' multilateral approaches to the misinformation problem were crafted for the small minority of people who strongly believe in or are inclined to believe in demonstrably false, or potentially harmful information. Framing the problem in this way indirectly assumes the binary split among individuals; i.e., you're either a neo-Nazi propagandist, or you're just vulnerable to their deception. In reality, there exists a considerable grey area within media ecosystems that is not necessarily represented by these internalization approaches. Largely speaking, a lot of people who are exposed to mis- and disinformation will not be affected; as Dr. Nyhan puts it, we must "focus more on how platforms can enable an extremist minority to foment harm and not on how the average person might be brainwashed by a piece of content they viewed a few times"

(Ovide, 2021). As such, the focal point of internalization approaches should be more directed toward the fringe *subgroups* who consume and perpetuate the largest amounts of harmful content. In other words, the people or groups that do not reach the largest majority of people, but could do a lot of harm to the people they do reach; i.e., the Twitter "superusers" from the Green et al. (2021) study.

To recognize there is no one-size-fits-all solution to misinformation is to appreciate the various combinations of proactive and reactive solutions. Steps in the right direction are representative of a time when there is an explicit motivation among politicians and the public alike to be well-informed, and when there are strong political, social, and professional incentives to do so (Lewandowsky et al., 2017). But above all, we, as a society, need to first learn how to ameliorate our deep polarization and be more open to opposing viewpoints outside of our digital echo chambers. Doing so necessitates setting aside our intractable oppositions in the interest of the common good. The misinformation problem is not about anti-conservative bias (Barrett and Sims, 2021) and it's not about the censorship of free speech (Denicola, 1979), and any argument placing either of these accusations at the forefront fails to understand the problem in its entirety. Former senator Daniel Patrick Moynihan frames it perfectly: "everyone is entitled to his own opinion, but not to his own facts."[71]

Looking forward, we must use the tools outlined here to pave the way forward. For platforms, this means emphasizing transparency, oversight, trust, and explainability in interventions, expanding English-centric approaches to other world

[71] See: *Daniel Patrick Moynihan: A Portrait in Letters of an American Visionary* (2010).

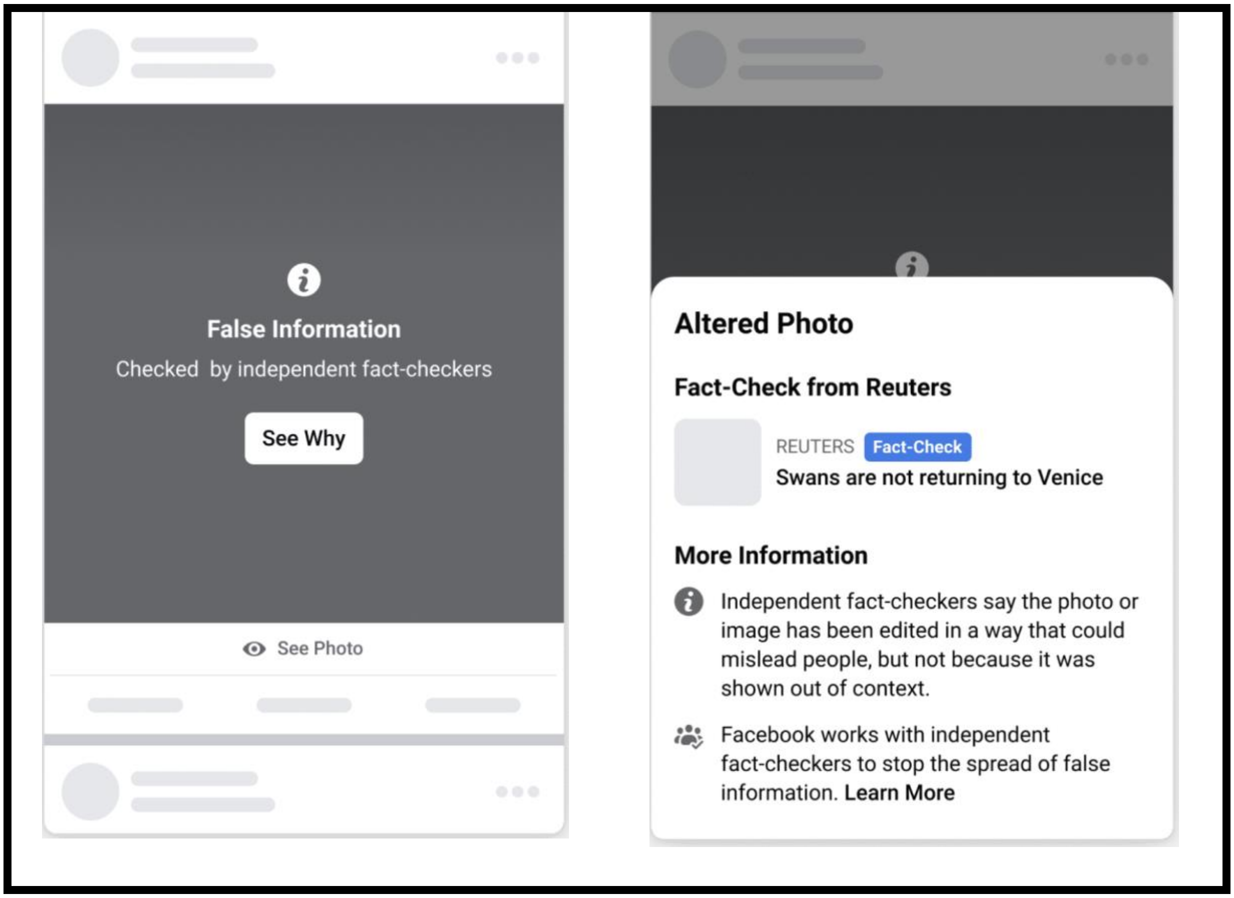
languages, and continuing to iterate these processes through community-driven feedback. For individuals, this means using what we know to inform what we ought to know – which could be as straightforward as simply knowing the differences between mis- and disinformation. In April 2022, when President Biden announced the creation of the 'Disinformation Governance Board' within the Department of Homeland Security to combat Russian disinformation and irregular migration, the polarization of the American polity was on full display. Conservative pundits and leaders alike sparked outrage, calling the board a "Ministry of Truth," a claim with no basis in fact. Knowing that *dis*information relates to unambiguous, propagandist dissemination, it is clear this governance board has nothing to do with monitoring domestic activity concerning *mis*information. Thus, fears of an Orwellian-style takeover are entirely unsubstantiated, given that matters within the DHS are centered around protecting national security from foreign interference – not undermining it.

If every step forward is met with a polarizing step backward, we will ultimately get nowhere. What makes the allure of democracy so enticing is the existence of opposing ideologies – such polarization should effectually make democracy stronger, not weaker. President Obama is right: it's up to us to determine what we prioritize and what values we define ourselves by. Ultimately, people define platforms, platforms don't define people. We have the toolbox – now let's use it like our democracy, our health, and our children's children depend on it.

8 Appendices

8.1 Appendix Figures

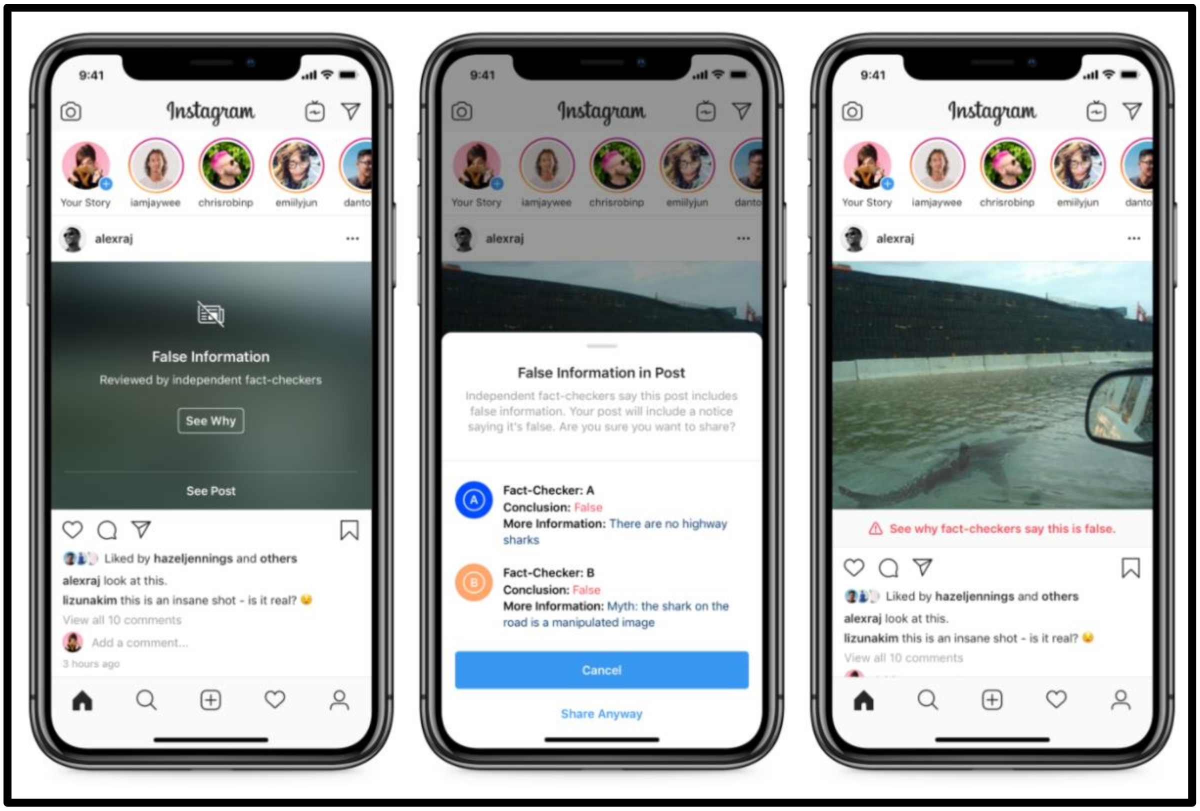
Figure 1: Notice on Facebook directing users to a third-party fact-checker (Meta, 2022)



Independent fact-checkers say the photo or image has been edited in a way that could mislead people, but not because it was shown out of context.

Facebook works with independent fact-checkers to stop the spread of false information. Learn More

Figure 2: Instagram credibility and context labels (Instagram, 2019)



Liked by **hazeljennings** and **others**

alexraj look at this.

lizunakim this is an insane shot - is it real? 😮

View all 10 comments

Add a comment...

3 hours ago

False Information in Post

Independent fact-checkers say this post includes false information. Your post will include a notice saying it's false. Are you sure you want to share?

Fact-Checker: A
Conclusion: False
More Information: There are no highway sharks

Fact-Checker: B
Conclusion: False
More Information: Myth: the shark on the road is a manipulated image

⚠ See why fact-checkers say this is false.

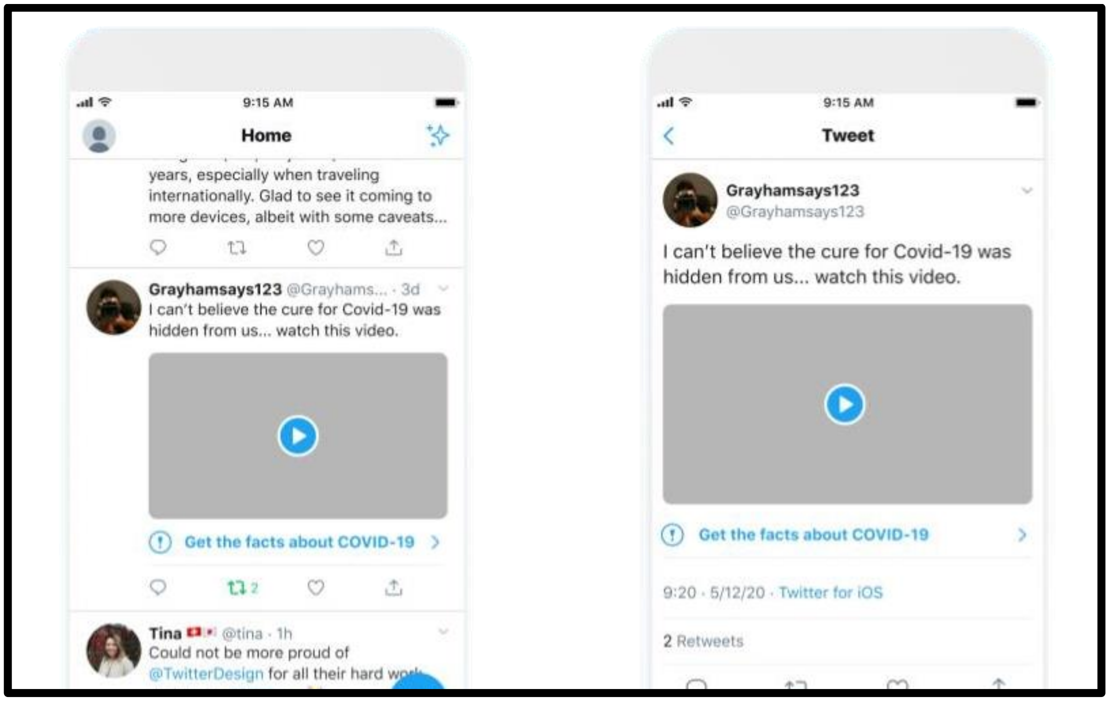
Liked by **hazeljennings** and **others**

alexraj look at this.

lizunakim this is an insane shot - is it real? 😮

View all 10 comments

Figure 3: Twitter contextual labels (Twitter, 2020)



years, especially when traveling internationally. Glad to see it coming to more devices, albeit with some caveats...

Grayhamsays123 @Grayhams... · 3d
I can't believe the cure for Covid-19 was hidden from us... watch this video.

Get the facts about COVID-19

Tina @tina · 1h
Could not be more proud of @TwitterDesign for all their hard work

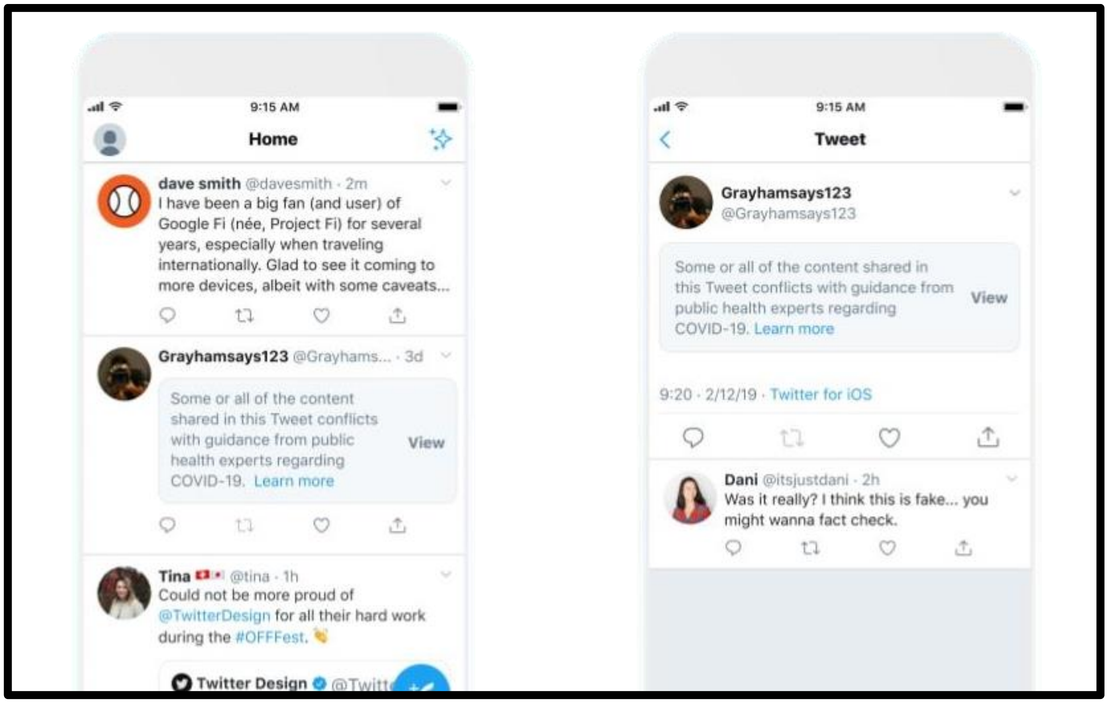
Grayhamsays123
@Grayhamsays123
I can't believe the cure for Covid-19 was hidden from us... watch this video.

Get the facts about COVID-19

9:20 · 5/12/20 · Twitter for iOS

2 Retweets

Figure 4: Twitter warning labels (Twitter, 2020)



dave smith @davesmith · 2m
I have been a big fan (and user) of Google Fi (née, Project Fi) for several years, especially when traveling internationally. Glad to see it coming to more devices, albeit with some caveats...

Grayhamsays123 @Grayhams... · 3d
Some or all of the content shared in this Tweet conflicts with guidance from public health experts regarding COVID-19. Learn more
View

Tina @tina · 1h
Could not be more proud of @TwitterDesign for all their hard work during the #OFFFest. 👏

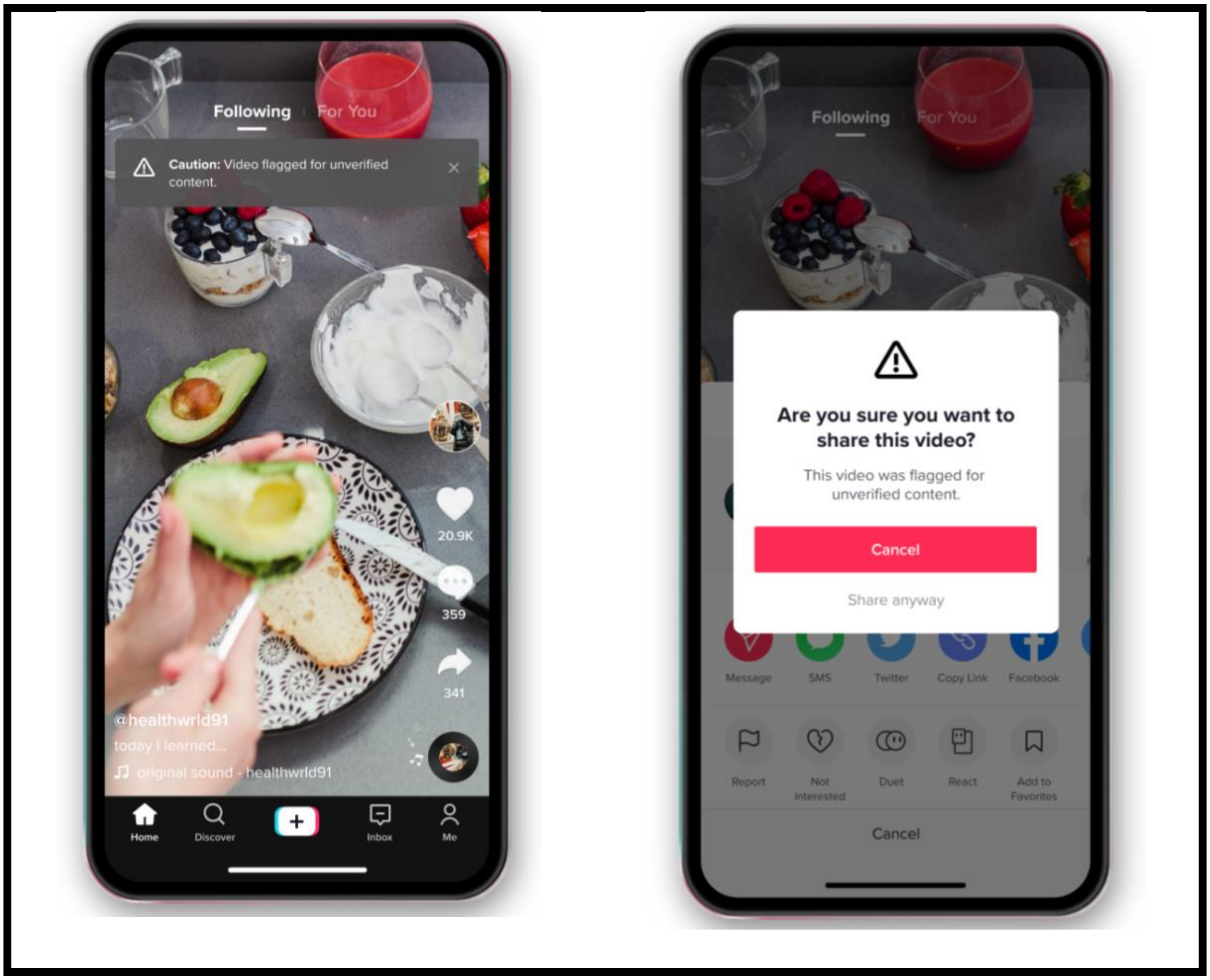
Twitter Design @Twitt

Grayhamsays123
@Grayhamsays123
Some or all of the content shared in this Tweet conflicts with guidance from public health experts regarding COVID-19. Learn more
View

9:20 · 2/12/19 · Twitter for iOS

Dani @itsjustdani · 2h
Was it really? I think this is fake... you might wanna fact check.

Figure 5: TikTok warning labels and prompt before sharing unverified content (TikTok, 2021)



Caution: Video flagged for unverified content.

Are you sure you want to share this video?

This video was flagged for unverified content.

Figure 6: Share of EU disinformation domains served and revenues paid, by company (GDI, 2020)

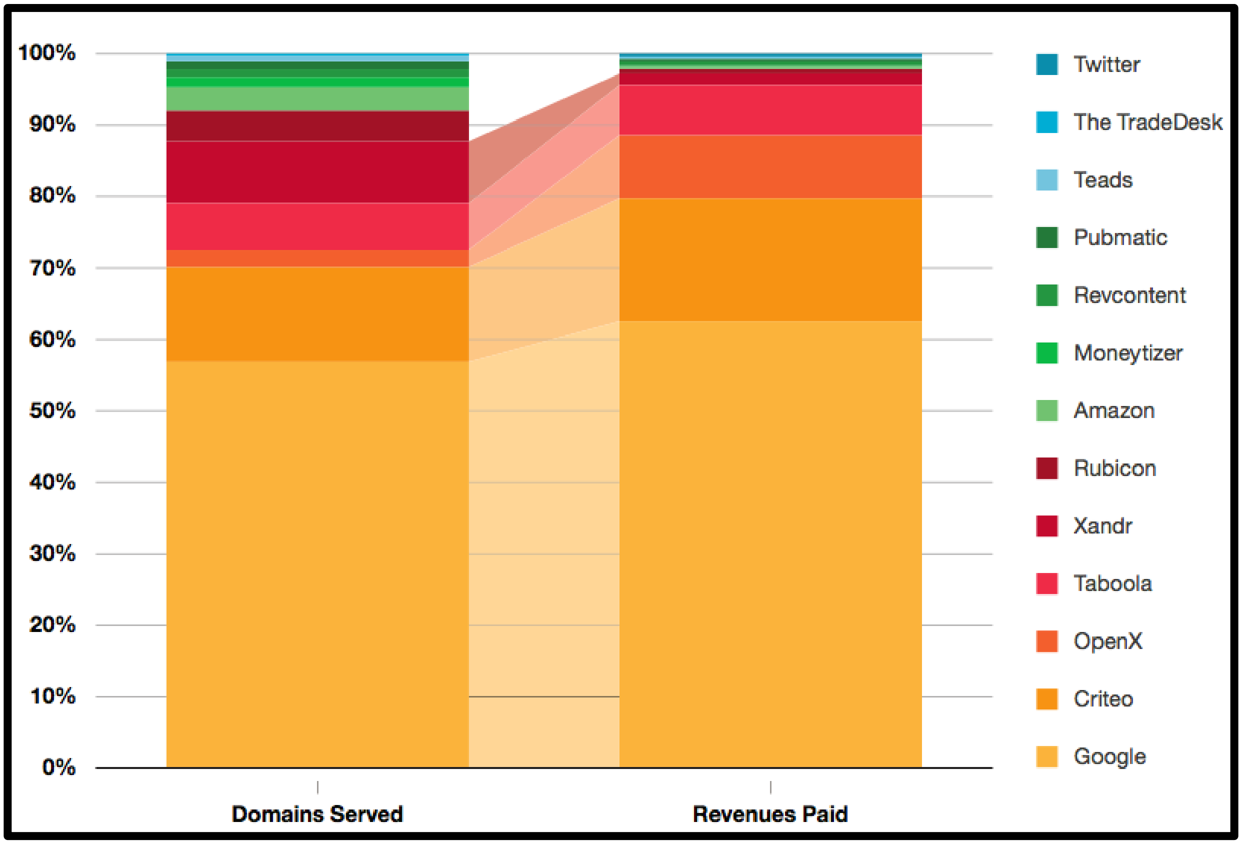
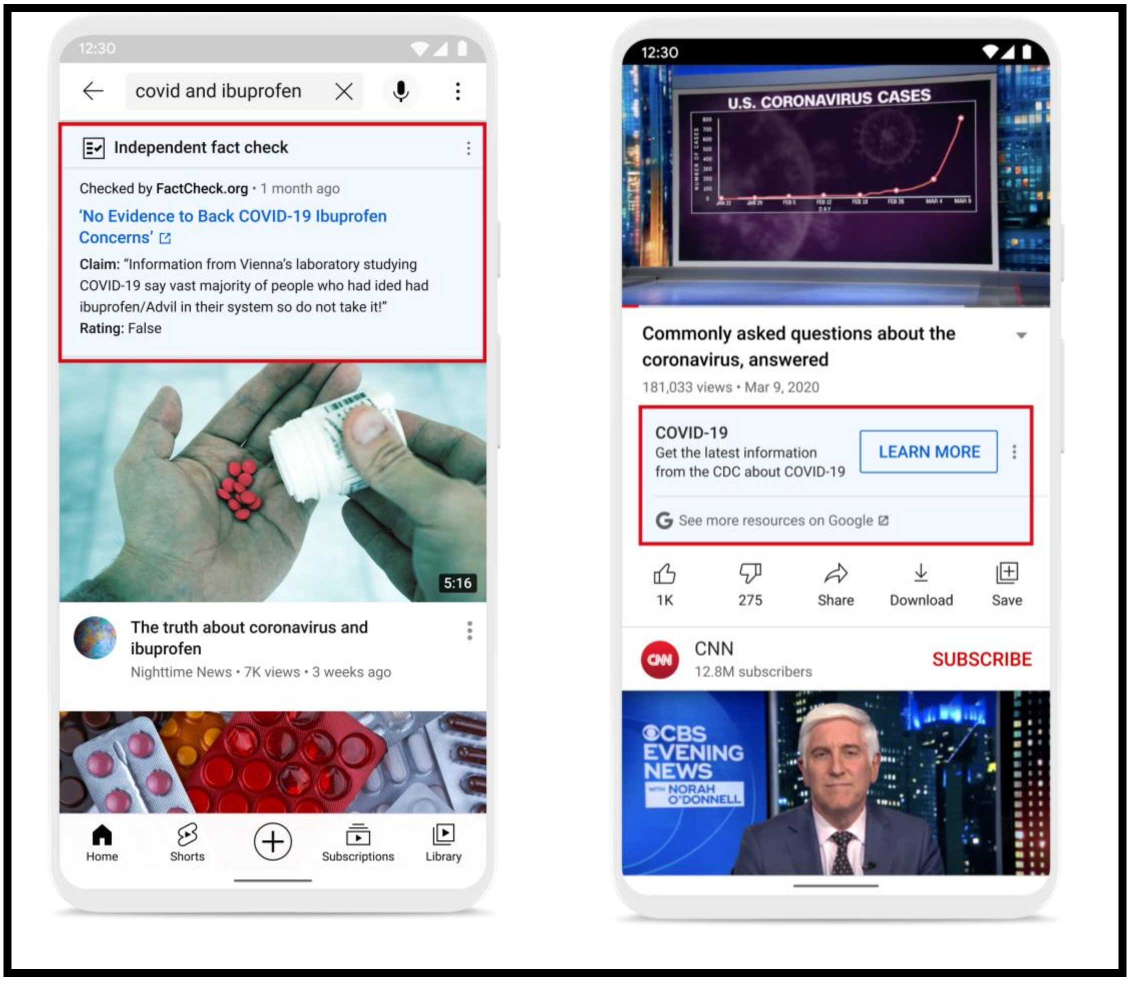


Figure 7: COVID-19 fact-check information panels on YouTube (YouTube, 2022)



12:30

covid and ibuprofen

Independent fact check

Checked by **FactCheck.org** • 1 month ago

'No Evidence to Back COVID-19 Ibuprofen Concerns'

Claim: "Information from Vienna's laboratory studying COVID-19 say vast majority of people who had ided had ibuprofen/Advil in their system so do not take it!"

Rating: False

5:16

The truth about coronavirus and ibuprofen

Nighttime News • 7K views • 3 weeks ago

12:30

Commonly asked questions about the coronavirus, answered

181,033 views • Mar 9, 2020

COVID-19

Get the latest information from the CDC about COVID-19

LEARN MORE

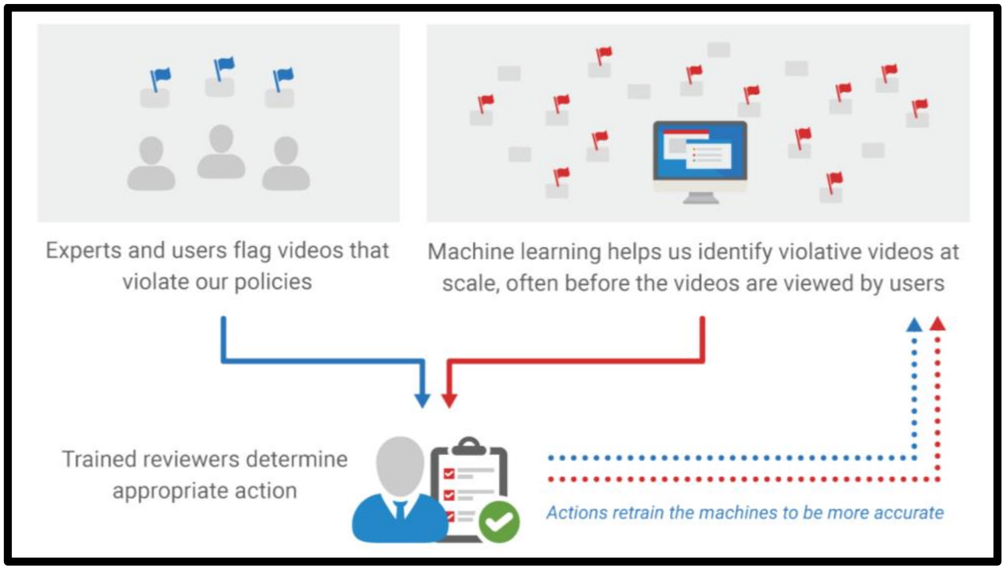
See more resources on Google

CNN

12.8M subscribers

SUBSCRIBE

Figure 8: Machine learning review system example on YouTube (Google, 2019)



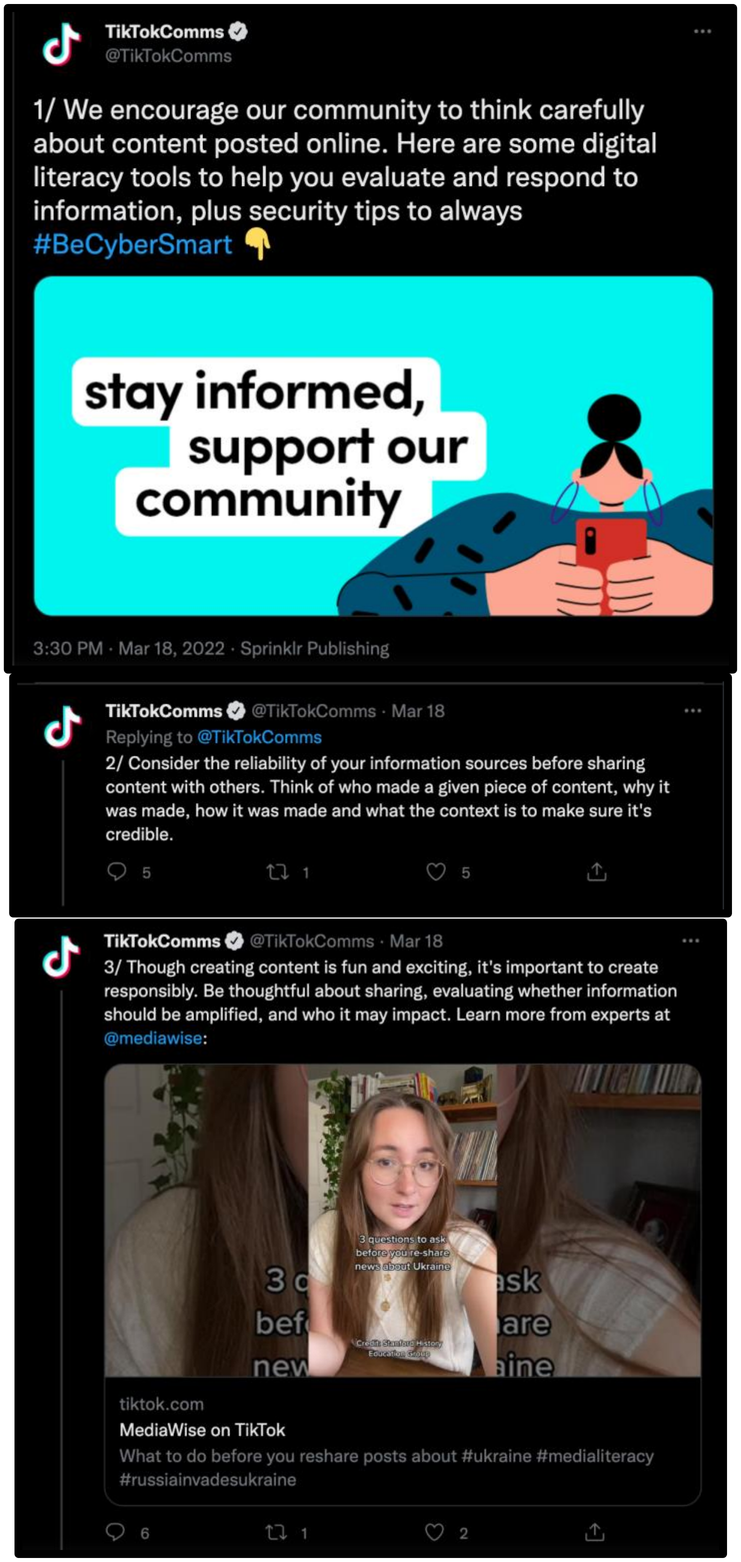
Experts and users flag videos that violate our policies

Machine learning helps us identify violative videos at scale, often before the videos are viewed by users

Trained reviewers determine appropriate action

Actions retrain the machines to be more accurate

Figure 9: TikTok digital literacy tools (TikTok, 2022)



TikTokComms
@TikTokComms

1/ We encourage our community to think carefully about content posted online. Here are some digital literacy tools to help you evaluate and respond to information, plus security tips to always #BeCyberSmart 👇

3:30 PM · Mar 18, 2022 · Sprinklr Publishing

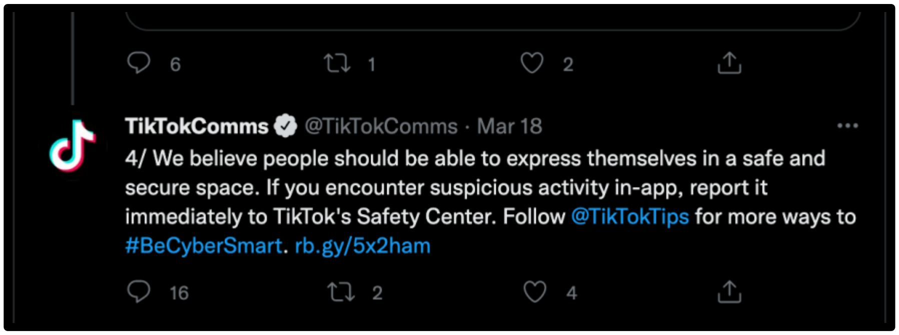
TikTokComms @TikTokComms · Mar 18
Replying to @TikTokComms

2/ Consider the reliability of your information sources before sharing content with others. Think of who made a given piece of content, why it was made, how it was made and what the context is to make sure it's credible.

TikTokComms @TikTokComms · Mar 18

3/ Though creating content is fun and exciting, it's important to create responsibly. Be thoughtful about sharing, evaluating whether information should be amplified, and who it may impact. Learn more from experts at @mediawise:

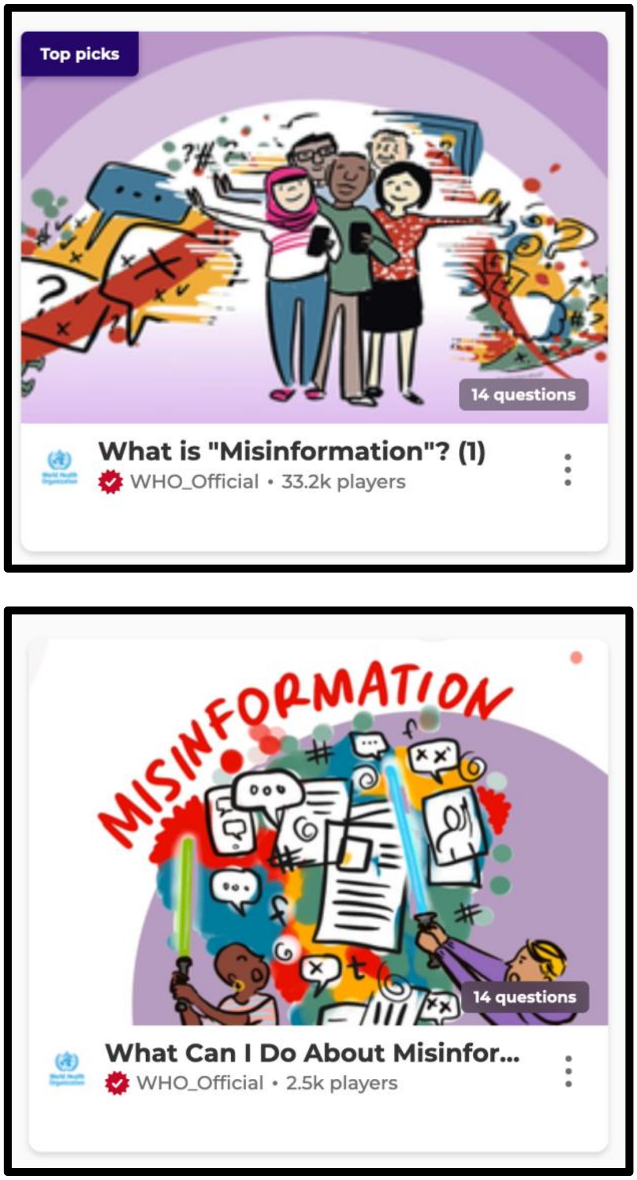
tiktok.com
MediaWise on TikTok
What to do before you reshare posts about #ukraine #medialiteracy #russiainvadesukraine



TikTokComms @TikTokComms · Mar 18

4/ We believe people should be able to express themselves in a safe and secure space. If you encounter suspicious activity in-app, report it immediately to TikTok's Safety Center. Follow @TikTokTips for more ways to #BeCyberSmart. rb.gy/5x2ham

Figure 10: Misinformation games and quizzes on Kahoot (Kahoot, 2022)



What is "Misinformation"? (1)

WHO_Official • 33.2k players

What Can I Do About Misinfor...

WHO_Official • 2.5k players

8.2 Appendix Tables

Table 1: Information Summary Characteristics

Characteristics	Information	Misinformation	Disinformation
Factual	Yes	Yes and no[72]	Yes and no
Complete	Yes and no	Yes and no	Yes and no
Intent	Yes and no	No	Yes
Informative	Yes	Yes	Yes
Deceptive	No	No	Yes

Table 2: Forms of Misinformation Interventions by Platforms on Individual Posts

Type	Definition	Example
Credibility label	Labels with attachments to authoritative sources or third-party fact-checkers.	Instagram and Facebook apply labels to posts based on fact-checker ratings such as "false information" (Facebook, 2021).
Contextual label	Information that provides additional context that the content of the user-generated post does not provide.	TikTok will detects and tag videos with words or hashtags related to the COVID-19 vaccine with the message 'Learn more about COVID-19 vaccines' and links to a COVID-19 "information hub" (TikTok, 2020).
Removal	The temporary or permanent removal of a post from a platform feed.	YouTube removed a COVID-19 conspiracy theory video ("plandemic") in May 2020.
Downranking	Reducing the number of times a post appears in other users' social media feeds.	Platforms downranking exaggerated or false health claims pertaining to COVID-19 treatments.

[72] Note: "Yes and no" depending on time, place, and context.

9 References

AACAP. “Social Media and Teens.” Accessed April 28, 2022. https://www.aacap.org/AACAP/Families_and_Youth/Facts_for_Families/FFF-Guide/Social-Media-and-Teens-100.aspx.

Abrahams, Alexei, and Gabrielle Lim. “Repress/Redress: What the ‘War on Terror’ Can Teach Us about Fighting Misinformation.” *Harvard Kennedy School Misinformation Review*, July 22, 2020. https://doi.org/10.37016/mr-2020-032.

Allcott, Hunt, and Matthew Gentzkow. "Social media and fake news in the 2016 election." *Journal of economic perspectives* 31, no. 2 (2017): 211-36.

Meta. “An Update on Our Work to Keep People Informed and Limit Misinformation About COVID-19,” April 16, 2020. https://about.fb.com/news/2020/04/covid-19-misinfo-update/.

Andrews, Phoenix. “Social Media Futures: What Is Brigading?” Tony Blair Institute for Global Change, March 2021. https://institute.global/policy/social-media-futures-what-brigading.

Ardia, David. “Free Speech Savior or Shield for Scoundrels: An Empirical Study of Intermediary Immunity under Section 230 of the Communications Decency Act.” *Loyola L.A. Law Review* 43, no. 2 (2010): 373,410.

Barrett, Paul, and Sims, Grant. NYU Stern Center for Business and Human Rights. “Tech - Bias Report 2021.” Accessed May 1, 2022. https://bhr.stern.nyu.edu/bias-report-release-page.

Basu, Tanya. “How to Talk to Kids and Teens about Misinformation | MIT Technology Review.” Accessed April 28, 2022. https://www.technologyreview.com/2020/11/02/1011528/election-2020-how-to-talk-to-kids-and-teens-about-misinformation/.

Bednar, Peter, and Christine Welch. *Bias Misinformation and the Paradox of Neutrality*, 2008. https://doi.org/10.28945/3277.

Brandom, Russell. “Trump Calls for Last-Minute 230 Repeal as Part of Defense Spending Bill.” The Verge, December 2, 2020.

https://www.theverge.com/2020/12/2/22037118/trump-section-230-repeal-ndaa-rider-facebook-twitter-moderation.

Bulger, Monica, and Patrick Davison. "The promises, challenges, and futures of media literacy." *Journal of Media Literacy Education* 10, no. 1 (2018): 1-21.

Bursztyn, Leonardo, Aakaash Rao, Christopher P. Roth, and David H. Yanagizawa-Drott. *Misinformation during a pandemic*. No. w27417. National Bureau of Economic Research, 2020.

Bursztyn, Leonardo, Georgy Egorov, Ruben Enikolopov, and Maria Petrova. *Social media and xenophobia: evidence from Russia*. No. w26567. National Bureau of Economic Research, 2019.

Byrd, Brian, and Smyser, Joseph. "Lies, Bots, and Coronavirus: Misinformation's Deadly Impact on Health," *Grantmakers In Health,* July 17, 2020. https://www.gih.org/views-from-the-field/lies-bots-and-coronavirus-misinformations-deadly-impact-on-health/.

Cannon, Robert. "The Legislative History of Senator Exon's Communications Decency Act: Regulating Barbarians on the Information Superhighway." *Federal Communications Law Journal* 49, no. 1 (November 1996): 52–94.

Card, David, and Gordon B. Dahl. "Family violence and football: The effect of unexpected emotional cues on violent behavior." *The quarterly journal of economics* 126, no. 1 (2011): 103-143.

Cattich, Ryan. "The Digital Pandemic." *Elements*, *17* (1), 21-32 (2020). https://doi.org/10.6017/eurj.v17i1.14915.

Chiang, Chun-Fang, and Brian Knight. "Media bias and influence: Evidence from newspaper endorsements." *The Review of economic studies* 78, no. 3 (2011):795-820.

TikTok. "Combating Misinformation and Election Interference on TikTok," August 16, 2019. https://newsroom.tiktok.com/en-us/combating-misinformation-and-election-interference-on-tiktok.

Instagram. “Combatting Misinformation on Instagram.” Accessed March 31, 2022. https://about.instagram.com/blog/announcements/combatting-misinformation-on-instagram.

Communications Decency Act, 47 U.S. § 230 (2018).

Reddit. “Content Policy.” Accessed April 22, 2022. https://www.redditinc.com/policies/content-policy.

Facebook. “Coronavirus (COVID-19) Information Center.” Accessed March 29, 2022. https://www.facebook.com/coronavirus_info.

FBK. “COVID-19 and Fake News in the Social Media.” Accessed April 24, 2022. https://www.fbk.eu/en/press-releases/covid-19-and-fake-news-in-the-social-media/.

Twitter. “COVID-19 Misleading Information Policy.” Accessed April 7, 2022. https://help.twitter.com/en/rules-and-policies/medical-misinformation-policy.

Dave, Aditya, Ioannis Vasileios Chremos, and Andreas A. Malikopoulos. “Social Media and Misleading Information in a Democracy: A Mechanism Design Approach.” *IEEE Transactions on Automatic Control*, 2021, 1–1. https://doi.org/10.1109/TAC.2021.3087466.

DellaVigna, Stefano, and Eliana La Ferrara. "Economic and social impacts of the media." In *Handbook of media economics*, vol. 1, pp. 723-768. North-Holland,2015.

Denicola, Robert C. “Copyright and Free Speech: Constitutional Limitations on the Protection of Expression.” *California Law Review* 67, no. 2 (1979): 283–316.

Doering, Jodi. “Twitter Post,” November 14, 2020. https://twitter.com/JodiOrth/status/1327771329555292162.

Eggers, Andrew C., Haritz Garro, and Justin Grimmer. “No Evidence for Systematic Voter Fraud: A Guide to Statistical Claims about the 2020 Election.” *Proceedings of the National Academy of Sciences* 118, no. 45 (November 9, 2021): e2103619118. https://doi.org/10.1073/pnas.2103619118.

European Parliament. “Understanding propaganda and disinformation” Accessed March 25, 2022.

https://www.europarl.europa.eu/RegData/etudes/ATAG/2015/571332/EPRS_ATA(2015)571332_EN.pdf.

Evon, Dan. “Is This ‘Ghost of Kyiv’ Video Real? | Snopes.Com.” Accessed April 10, 2022. https://www.snopes.com/fact-check/is-this-ghost-of-kyiv-video-real/.

Farrell, Henry John, and Bruce Schneier. “Common-Knowledge Attacks on Democracy.” *SSRN Electronic Journal*, 2018. https://doi.org/10.2139/ssrn.3273111.

Cybersecurity for Democracy. “Far Right News Sources on Facebook More Engaging.” *Cybersecurity for Democracy* (blog), March 4, 2021. https://medium.com/cybersecurity-for-democracy/far-right-news-sources-on-facebook-more-engaging-e04a01efae90.

Feiner, Lauren. “Reddit Users Are the Least Valuable of Any Social Network.” CNBC, February 11, 2019. https://www.cnbc.com/2019/02/11/reddit-users-are-the-least-valuable-of-any-social-network.html.

Fischer, Sara. “Global Advertising Industry Expected to Hit $1 Trillion by 2025.” Axios, December 7, 2021. https://www.axios.com/advertising-industry-revenue-9147f591-3e74-48bb-ab9f-c352f7283a48.html.

Fischer, Zachary Basu, Sara. “Russian Disinformation Frenzy Seeds Groundwork for Ukraine Invasion.” Axios. Accessed February 21, 2022. https://www.axios.com/russian-disinformation-frenzy-seeds-groundwork-for-ukraine-invasion-5d819df4-8f59-4320-ae05-5517ecae81d1.html.

Fisher, Natascha A. Karlova, Karen E. “A Social Diffusion Model of Misinformation and Disinformation for Understanding Human Information Behaviour.” Text. Professor T.D. Wilson, March 15, 2013. http://informationr.net/ir/18-1/paper573.html#.Yj5fYxDMLv0.

Fox, Christopher John. *Information and Misinformation: An Investigation of the Notions of Information, Misinformation, Informing, and Misinforming*. Greenwood Publishing Group, 1983.

Gawande, Atul. “Twitter Post,” November 16, 2020. https://twitter.com/Atul_Gawande/status/1328409592959823875.

Goldman, Eric. “Dear President Biden: You Should Save, Not Revoke, Section 230.” *Bulletin of the Atomic Scientists* 77, no. 1 (January 2, 2021): 36–37. https://doi.org/10.1080/00963402.2020.1859863.

Goldman, Eric. “The Complicated Story of FOSTA and Section 230.” *First Amendment Law Review* 17 (2019 2018): 279–93.

Goldman, Eric. “The Ten Most Important Section 230 Rulings.” *Tulane Journal of Technology and Intellectual Property* 20 (2017): 1–10.

Google. “Google Publisher Policies - Google AdSense Help.” Accessed April 14, 2022. https://support.google.com/adsense/answer/10502938?hl=en&visit_id=637717920975924126-977395725&rd=1.

Green, Jon, William Hobbs, Stefan McCabe, and David Lazer. “Despair or Defiance? Turnout and Online Engagement with Election Misinformation,” December 14, 2021.

Grice, Herbert P. "Logic and conversation." In *Speech acts*, pp. 41-58. Brill, 1975.

Halon, Yael. “Zuckerberg Knocks Twitter for Fact-Checking Trump, Says Private Companies Shouldn’t Be ‘the Arbiter of Truth.’” Text. Article. Fox News. Fox News, May 27, 2020. https://www.foxnews.com/media/facebook-mark-zuckerberg-twitter-fact-checking-trump.

Hao, Karen. “How Facebook and Google Fund Global Misinformation.” *MIT Technology Review*. Accessed February 7, 2022. https://www.technologyreview.com/2021/11/20/1039076/facebook-google-disinformation-clickbait/.

Hao, Karen. “How Facebook Got Addicted to Spreading Misinformation.” *MIT Technology Review*. Accessed April 27, 2022. https://www.technologyreview.com/2021/03/11/1020600/facebook-responsible-ai-misinformation/.

Hasson, Uri, Joseph P. Simmons, and Alexander Todorov. "Believe it or not: On the possibility of suspending belief." *Psychological science* 16, no. 7 (2005): 566-571.

Hernández-Huerta, Victor, and Francisco Cantú. "Public Distrust in Disputed Elections: Evidence from Latin America." *British Journal of Political Science*, November 8, 2021, 1–8. https://doi.org/10.1017/S0007123421000399.

Hossain, Tamanna, Robert L. Logan IV, Arjuna Ugarte, Yoshitomo Matsubara, Sean Young, and Sameer Singh. "COVIDLies: Detecting COVID-19 Misinformation on Social Media." In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*. Online: Association for Computational Linguistics, 2020. https://doi.org/10.18653/v1/2020.nlpcovid19-2.11.

Avaaz. "How Facebook Can Flatten the Curve of the Coronavirus Infodemic." Accessed March 29, 2022. https://secure.avaaz.org/campaign/en/facebook_coronavirus_misinformation/.

Google. "How Fact Checkers and Google.Org Are Fighting Misinformation," March 31, 2021. https://blog.google/outreach-initiatives/google-org/fullfact-and-google-fight-misinformation/.

Google. "Impact Report - Google News Initiative." Accessed April 15, 2022. http://localhost/impact2021/.

Quote Investigator. "In a Time of Universal Deceit — Telling the Truth Is a Revolutionary Act – Quote Investigator." Accessed February 21, 2022. https://quoteinvestigator.com/2013/02/24/truth-revolutionary/.

Ingram, Mathew. "The Myth of Social Media Anti-Conservative Bias Refuses to Die." *Columbia Journalism Review*. Accessed March 20, 2022. https://www.cjr.org/the_media_today/platform-bias.php.

Twitter. "Introducing Birdwatch, a Community-Based Approach to Misinformation." Accessed April 7, 2022. https://blog.twitter.com/en_us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation.

The Economist. "Is Investing in Twitter a Meme Too Far for Elon Musk?," April 9, 2022. https://www.economist.com/business/is-investing-in-twitter-a-meme-too-far-for-elon-musk/21808586.

Jensen, Robert, and Emily Oster. "The power of TV: Cable television and women's status in India." *The Quarterly Journal of Economics* 124, no. 3 (2009): 1057-1094.

Jiménez Durán, Rafael. "The Economics of Content Moderation: Theory and Experimental Evidence from Hate Speech on Twitter." *Available at SSRN* (2022).

Kaplan, Richard L. *Politics and the American Press: The Rise of Objectivity, 1865-1920*. Cambridge University Press, 2002.

Karlova, Natascha A., and Jin Ha Lee. "Notes from the Underground City of Disinformation: A Conceptual Investigation." *Proceedings of the American Society for Information Science and Technology* 48, no. 1 (2011): 1–9.

Kasperkevic, Jana. "Nobel Laureate Paul Romer on How to Curb Big Tech's Power | University of Chicago News." Accessed April 28, 2022. https://news.uchicago.edu/story/nobel-laureate-paul-romer-how-curb-big-techs-power.

Klepper, David. "Facebook Clarifies Zuckerberg Remarks on False Political Ads." ABC News. Accessed April 7, 2022. https://abcnews.go.com/Politics/wireStory/facebook-clarifies-zuckerberg-remarks-false-political-ads-66513056.

Kearney, Melissa S., and Phillip B. Levine. "Media influences on social outcomes: The impact of MTV's 16 and pregnant on teen childbearing." *American Economic Review* 105, no. 12 (2015): 3597-3632.

Kuklinski, James H., Paul J. Quirk, Jennifer Jerit, David Schwieder, and Robert F. Rich. "Misinformation and the currency of democratic citizenship." *The Journal of Politics* 62, no. 3 (2000): 790-816.

La Ferrara, Eliana. "Mass media and social change: Can we use television to fight poverty?." *Journal of the European Economic Association* 14, no. 4 (2016): 791-827.

Lang, Kurt, and Gladys Engel Lang. *Television and Politics*. New York: Routledge, 2017. https://doi.org/10.4324/9781351306089.

Lazarus, Richard S. *Emotion and Adaptation*. Oxford University Press, 1991.

Lee, Jung Jae, Kyung-Ah Kang, Man Ping Wang, Sheng Zhi Zhao, Janet Yuen Ha Wong, Siobhan O'Connor, Sook Ching Yang, and Sunhwa Shin. "Associations Between COVID-19 Misinformation Exposure and Belief With COVID-19 Knowledge and Preventive Behaviors: Cross-Sectional Online Study." *Journal of Medical Internet Research* 22, no. 11 (November 13, 2020): e22205. https://doi.org/10.2196/22205.

Long, Elisa F., M. Keith Chen, and Ryne Rohla. "Political storms: Emergent partisan skepticism of hurricane risks." *Science advances* 6, no. 37 (2020): eabb7906.

Pew Research. "Less-Educated Americans More Inclined to See Some Truth in Conspiracy Theory That COVID-19 Was Planned." *Pew Research Center* (blog). Accessed April 28, 2022. https://www.pewresearch.org/wp-content/uploads/2020/07/FT_20.07.15_Conspiracies_feature-1.png.

Lewandowsky, Stephan, Ullrich K. H. Ecker, and John Cook. "Beyond Misinformation: Understanding and Coping with the 'Post-Truth' Era." *Journal of Applied Research in Memory and Cognition* 6, no. 4 (December 1, 2017): 353–69. https://doi.org/10.1016/j.jarmac.2017.07.008.

Lewandowsky, Stephan, Ullrich K. H. Ecker, Colleen M. Seifert, Norbert Schwarz, and John Cook. "Misinformation and Its Correction: Continued Influence and Successful Debiasing." *Psychological Science in the Public Interest* 13, no. 3 (December 1, 2012): 106–31. https://doi.org/10.1177/1529100612451018.

Lorenz, Taylor. "TikTok Stars Receive White House Briefing on Ukraine." *The Washington Post*. Accessed April 11, 2022.

Lotty, Alexandra. "Apps Too: Modifying Interactive Computer Service Provider Immunity under Section 230 of the Communications Decency Act in the Wake of 'Me Too.'" *Southern California Law Review* 93, no. 4 (Spring 2020): 855–921.

Luenberger, D. G. "New Optimality Principles for Economic Efficiency and Equilibrium." *Journal of Optimization Theory and Applications* 75, no. 2 (November 1, 1992): 221–64. https://doi.org/10.1007/BF00941466.

McDougall, Julian. “Media Literacy versus Fake News: Critical Thinking, Resilience and Civic Engagement.” *Media Studies* 10, no. 19 (October 21, 2019): 29–45.

Mellor, Sophie. “TikTok’s Algorithm Shows Users Fake News on Ukraine War.” *Fortune*. Accessed April 10, 2022. https://fortune.com/2022/03/21/tiktok-misinformation-ukraine/.

Merrill, Jeremy, and McCarthy, Ryan. “Trump Won Florida After Running a False Ad Tying Biden to Venezuelan Socialists.” *ProPublica*. Accessed April 20, 2022. https://www.propublica.org/article/trump-won-florida-after-running-a-false-ad-tying-biden-to-venezuelan-socialists.

Mickle, Tripp. “Google Revenue Surges as Online Advertising Market Thrives.” WSJ. Accessed April 14, 2022. https://www.wsj.com/articles/google-alphabet-googl-2q-earnings-report-2021-11627344309.

Miguel, Edward, and Michael Kremer. “Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities.” *Econometrica* 72, no. 1 (2004): 159–217. https://doi.org/10.1111/j.1468-0262.2004.00481.x.

Milmo, Dan. “YouTube Is Major Conduit of Fake News, Factcheckers Say.” *The Guardian*, January 12, 2022, sec. Technology. https://www.theguardian.com/technology/2022/jan/12/youtube-is-major-conduit-of-fake-news-factcheckers-say.

Müller, Karsten, and Carlo Schwarz. "Fanning the flames of hate: Social media and hate crime." *Journal of the European Economic Association* 19, no. 4 (2021): 2131-2167.

NewsGuard. “Misinformation Monitor: March 2022.” Accessed April 10, 2022. https://www.newsguardtech.com/misinformation-monitor/march-2022.

YouTube. “Misinformation Policies - YouTube Help.” Accessed April 20, 2022. https://support.google.com/youtube/answer/10834785?hl=en.

Mitchell, Charlie. “Inside Cybersecurity,” August 4, 2020. https://insidecybersecurity.com/share/11499.

Musk, Elon. “Twitter Post,” March 27, 2022. https://twitter.com/Atul_Gawande/status/1328409592959823875.

Naeem, Salman Bin, Rubina Bhatti, and Aqsa Khan. "An exploration of how fake news is taking over social media and putting public health at risk." *Health Information & Libraries Journal* 38, no. 2 (2021): 143-149.

Nilsen, Jennifer, Fagan, Kaylee, Dreyfuss, Emily, and Donovan, Joan. Media Manipulation Casebook. “TikTok, the War on Ukraine, and 10 Features That Make the App Vulnerable to Misinformation,” March 10, 2022. https://mediamanipulation.org/research/tiktok-war-ukraine-and-10-features-make-app-vulnerable-misinformation.

Statista. “Number of Social Media Users 2025.” Accessed March 14, 2022. https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/.

Nunziato, Dawn Carla. “Misinformation Mayhem: Social Media Platforms’ Efforts to Combat Medical and Political Misinformation.” *First Amendment Law Review* 19, no. 1 (2021 2020): 32–98.

Ognyanova, Katherine, David Lazer, Ronald E. Robertson, and Christo Wilson. “Misinformation in Action: Fake News Exposure Is Linked to Lower Trust in Media, Higher Trust in Government When Your Side Is in Power.” *Harvard Kennedy School Misinformation Review*, June 2, 2020. https://doi.org/10.37016/mr-2020-024.

Facebook. “Our Approach to Misinformation | Transparency Center.” Accessed March 29, 2022. https://transparency.fb.com/features/approach-to-misinformation/.

Ovide, Shira. “YouTube’s Ban on Misinformation.” *The New York Times*, October 5, 2021, sec. Technology. https://www.nytimes.com/2021/10/05/technology/youtube-misinformation.html.

Oxford Languages. “Oxford Word of the Year 2016.” Accessed March 14, 2022. https://languages.oup.com/word-of-the-year/2016/.

Paul, Kari. “‘Democracy Will Wither’: Barack Obama Outlines Perils of Unregulated Big Tech in Sweeping Speech.” *The Guardian*, April 21, 2022, sec. US news. https://www.theguardian.com/us-news/2022/apr/21/obama-stanford-speech-big-tech.

Paul, Kari. "TikTok Was 'Just a Dancing App'. Then the Ukraine War Started." *The Guardian*, March 20, 2022, sec. Technology. https://www.theguardian.com/technology/2022/mar/19/tiktok-ukraine-russia-war-disinformation.

Przeworski, Adam. "Minimalist Conception of Democracy: A Defense." *Defining Democracy*, n.d., 6.

Quinn, Emma K., Sajjad S. Fazel, and Cheryl E. Peters. "The Instagram Infodemic: Cobranding of Conspiracy Theories, Coronavirus Disease 2019 and Authority-Questioning Beliefs." *Cyberpsychology, Behavior, and Social Networking* 24, no. 8 (August 2021): 573–77. https://doi.org/10.1089/cyber.2020.0663.

Instagram. "Reducing the Spread of False Information on Instagram | Instagram Help Center." Accessed March 31, 2022. https://help.instagram.com/1735798276553028.

Reimann, Nicholas. "QAnon Supporters Pack Site Of JFK Assassination In Hopes JFK Jr. (And Maybe His Dad) Will Return." Forbes. Accessed February 21, 2022. https://www.forbes.com/sites/nicholasreimann/2021/11/22/qanon-supporters-pack-site-of-jfk-assassination-in-hopes-jfk-jr-and-maybe-his-dad-will-return/.

Richards, Abbie. "A Pro-Russia Propaganda Campaign Is Using over 180 TikTok Influencers to Promote the Invasion of Ukraine." Media Matters for America. Accessed April 11, 2022. https://www.mediamatters.org/tiktok/pro-russia-propaganda-campaign-using-over-180-tiktok-influencers-promote-invasion-ukraine.

Riker, William H., and Peter C. Ordeshook. "A Theory of the Calculus of Voting." *American Political Science Review* 62, no. 1 (March 1968): 25–42. https://doi.org/10.2307/1953324.

Rogers, Richard. "Research Note: The Scale of Facebook's Problem Depends upon How 'Fake News' Is Classified." *Harvard Kennedy School Misinformation Review*, October 26, 2020. https://doi.org/10.37016/mr-2020-43.

Romer, Paul. "Opinion | A Tax That Could Fix Big Tech." *The New York Times*, May 6, 2019, sec. Opinion. https://www.nytimes.com/2019/05/06/opinion/tax-facebook-google.html.

Rubin, Victoria L. "On Deception and Deception Detection: Content Analysis of Computer-Mediated Stated Beliefs." *Proceedings of the American Society for Information Science and Technology* 47, no. 1 (2010): 1–10. https://doi.org/10.1002/meet.14504701124.

Rushkoff, Douglas. *Team Human*. Ledizioni, 2020.

Saltz, Emily, Soubhik Barari, Claire Leibowicz, and Claire Wardle. "Misinformation Interventions Are Common, Divisive, and Poorly Understood." *Harvard Kennedy School Misinformation Review*, October 27, 2021. https://doi.org/10.37016/mr-2020-81.

Sanders, Michael. "The Problem with Pareto." Accessed March 24, 2022. https://www.kcl.ac.uk/news/the-problem-with-pareto.

Sanderson, Zeve, Megan A. Brown, Richard Bonneau, Jonathan Nagler, and Joshua A. Tucker. "Twitter Flagged Donald Trump's Tweets with Election Misinformation: They Continued to Spread Both on and off the Platform." *Harvard Kennedy School Misinformation Review*, August 24, 2021. https://doi.org/10.37016/mr-2020-77.

Sandler, Todd, and V. Kerry Smith. "Intertemporal and Intergenerational Pareto Efficiency." *Journal of Environmental Economics and Management* 2, no. 3 (February 1, 1976): 151–59. https://doi.org/10.1016/0095-0696(76)90030-9.

Schul, Yaacov, Ruth Mayo, and Eugene Burnstein. "The Value of Distrust." *Journal of Experimental Social Psychology* 44, no. 5 (September 1, 2008): 1293–1302. https://doi.org/10.1016/j.jesp.2008.05.003.

Schwarz, Norbert. "Judgment in a Social Context: Biases, Shortcomings, and the Logic of Conversation." In *Advances in Experimental Social Psychology*, edited by Mark P. Zanna, 26:123–62. Academic Press, 1994. https://doi.org/10.1016/S0065-2601(08)60153-7.

Seetharaman, Jeff Horwitz and Deepa. "Facebook Executives Shut Down Efforts to Make the Site Less Divisive." WSJ. Accessed April 27, 2022.

https://www.wsj.com/articles/facebook-knows-it-encourages-division-top-executives-nixed-solutions-11590507499.

Sloss, David. “Section 230 and the Duty to Prevent Mass Atrocities.” *Case Western Reserve Journal of International Law* 52, no. 1/2 (Spring 2020): 199–212.

Solon, Olivia, and Sam Levin. “How Google’s Search Algorithm Spreads False Information with a Rightwing Bias.” *The Guardian*, December 16, 2016, sec. Technology. https://www.theguardian.com/technology/2016/dec/16/google-autocomplete-rightwing-bias-algorithm-political-propaganda.

NewsGuard. “Special Report: How Some of the World’s Largest Brands Funded the Misinformation behind the Capitol Riot.” Accessed April 14, 2022. https://www.newsguardtech.com/special-reports/special-report-advertising-on-election-misinformation.

Sperber, Dan, and Deirdre Wilson. *Relevance*: *Communication and cognition*. Vol. 142. Cambridge, MA: Harvard University Press, 1986.

Stahl, Bernd Carsten. “On the Difference or Equality of Information, Misinformation, and Disinformation: A Critical Research Perspective.” *Informing Science Journal* 9 (2006): 83–96.

OHCHR. “Statement by Human Rights Council President at Glion Human Rights Dialogue on the Future of the Human Rights Council.” Accessed March 31, 2022. https://www.ohchr.org/en/statements/2019/05/statement-human-rights-council-president-glion-human-rights-dialogue-future.

Stevenson, Richard W. “Trying a Market Approach to Smog.” *The New York Times*. Accessed March 26, 2022. https://www.nytimes.com/1992/03/25/business/trying-a-market-approach-to-smog.html.

Taddicken, Monika, and Laura Wolff. “‘Fake News’ in Science Communication: Emotions and Strategies of Coping with Dissonance Online.” *Media and Communication* 8, no. 1 (March 18, 2020): 206–17.

GDI. “The Global Disinformation Index.” Accessed April 14, 2022. https://www.disinformationindex.org/.

Backlinko. “TikTok User Statistics (2022),” November 4, 2020. https://backlinko.com/tiktok-users.

Thompson, Alex. “Trump Deploys YouTube as His Secret Weapon in 2020.” *Politico.* Accessed April 21, 2022. https://www.politico.com/news/2020/09/06/trumpyoutube-election-comeback-408576.

Statista. “Top U.S. Mobile Social Apps by Session Length 2019.” Accessed April 11, 2022. https://www.statista.com/statistics/579411/top-us-social-networking-apps-ranked-by-session-length/.

Dailymotion. “Tucker Carlson Tonight 2-25-20 FULL - Breaking TRUMP February 25, 2020 - Video Dailymotion,” February 26, 2020. https://www.dailymotion.com/video/x7s8elj.

Verveer, Phillip. “Countering Negative Externalities in Digital Platforms,” October 7, 2019. https://shorensteincenter.org/countering-negative-externalities-in-digital-platforms/.

Yanagizawa-Drott, David. "Propaganda and conflict: Evidence from the Rwandan genocide." *The Quarterly Journal of Economics* 129, no. 4 (2014): 1947-1994.

YouTube. “YouTube Misinformation - How YouTube Works.” Accessed April 21, 2022. https://www.youtube.com/howyoutubeworks/our-commitments/fighting-misinformation/.