Disinformation, Stochastic Harm, and Costly Effort: A Principal-Agent Analysis of Regulating Social Media Platforms

by

Shehroze Khan

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

Department of Computing Science
University of Alberta

Abstract

The spread of disinformation on social media platforms is harmful to society. This harm may manifest as a gradual degradation of public discourse; but it can also take the form of sudden dramatic events such as the 2021 insurrection on Capitol Hill. The social media platforms themselves are in the best position to prevent the spread of disinformation, as they have the best access to relevant data and the expertise to use it. However, mitigating disinformation is costly, not only for implementing detection algorithms or employing manual effort, but also because moderating content impacts user engagement and thus potential advertising revenue. Since the costs of harmful content are borne by other entities, the platform will therefore have no incentive to exercise the socially-optimal level of effort.

A similar problem exists for the environmental regulation domain, where the costs of adverse events are not directly borne by a firm, the mitigation effort of a firm is not observable, and the causal link between a harmful consequence and a specific failure is difficult to prove. For environmental regulation, one solution is to perform costly monitoring to ensure that the firm takes adequate precautions according to a specified rule. However, a fixed rule for classifying disinformation becomes less effective over time, as bad actors can learn to sequentially and strategically bypass it.

In this thesis, we develop a formal model to capture incentives of social platforms relating to the control of online disinformation; our framework incorporates these important features of the disinformation prevention domain. Encoding our domain as a Markov decision process, we demonstrate that no penalty based on a static rule, no matter how large, can incentivize adequate effort. Penalties based on an adap-

tive rule can incentivize optimal effort, but counterintuitively, only if the regulator sufficiently *overreacts* to harmful events by requiring a greater-than-optimal level of effort. We discuss key implications of our formal results, highlight inherent challenges of regulating disinformation, and provide promising directions for future work.

Preface

Parts of this thesis have been accepted to the Cooperative AI Workshop at the Thirty-Fifth Conference on Neural Information Processing Systems (NeurIPS 2021), which is a non-archival venue. No part of this thesis has been previously published at an archival venue. However, we have submitted two additional papers based on this thesis to The Twenty-Third ACM Conference on Economics and Computation (EC'22) and also to the Proceedings of the National Academy of Sciences (PNAS).

The formal model, theoretical results and described analyses presented in this thesis are part of a collaborative research effort with my supervisor, Dr. James R. Wright.

To my family and my friends,

thank you for always believing in me.

. . .

Especially to my parents and to my sisters, Maheen and Meerub,

for your unconditional love, support and encouragement throughout my educational pursuits, and for always pushing me to become the best version of myself, thank you.

Acknowledgements

I would like to extend my deepest gratitude to my supervisor, Dr. James R. Wright, without whose support this thesis would not have come to fruition. James has played an integral part in my intellectual growth over these past couple of years, and also in my development as a young researcher. He has always encouraged me to go the extra mile with my research, pushed me to challenge my instinctive assumptions, and taught me indispensable methods of breaking down and solving complex problems. James' consistently productive feedback has not only helped me become a better writer, but also a clearer thinker; and his infectious enthusiasm for behavioral game theory is something I value immensely, as it has made for a very enjoyable journey of formalizing and executing on our ideas for this thesis. I am also grateful to James for his patience and understanding throughout all the uncertainty of the pandemic.

I must also thank members of the ABGT reading group for their insightful discussions during my presentations. Even outside of these discussions, it has been very comforting to have our small research group stay connected over these past two years, regardless of wherever we've been in the world. I'm also grateful for the friendship of my peers, especially Chris, Doug, Kristen, Cande, Rebecca and Sahir, all of who made my experience of moving to a completely new city much less daunting.

I would also like to acknowledge the generous funding and support I've received from the University of Alberta and Amii.

Finally, I don't even have words to express my thanks to my family, my parents and my sisters. Mama, Baba, Maheen, and Meerub, without your unwavering support, love and prayers, I would not have been able to accomplish anything.

Table of Contents

Chapter 1

Introduction

Contemporary web and social media platforms provide a ripe ground for the spread of false news, hoaxes, and disinformation [1]. Compounding the problem, social platforms' business models often conflict with efforts that can mitigate these problems. Facebook, for instance, uses machine learning models to maximize user engagement. In doing so, however, these models also favor content that is toxic and filled with conspiracy, lies, and misleading, divisive information [2–4].

We use *disinformation* to refer to all such toxic content, including all kinds of false and fabricated news posing as truth, created with the intention to mislead [1]. The unmitigated spread of disinformation is harmful to society. The harm can be direct physical or emotional distress to an individual; it may also manifest as a negative externality affecting public discourse, or social welfare. Examples include the undermining of public health response due to Covid-19 false rumors [5, 6], disease outbreaks via anti-vaccination propaganda [7], violent conspiracy movements surrounding the 2020 US presidential elections [8, 9], and horrific incidents such as the Pizzagate shooting [10], or ethnic violence in Myanmar [11].

The costs of these rare and dramatic events are borne exclusively by society, rather than the social platforms themselves. Furthermore, these events are inherently stochastic as it is impossible to predict with certainty that a given collection of content will cause a specific harm. And mitigating disinformation is costly: Fil-

tering, demoting or assigning warning labels to associated content entails both the direct costs of implementing classification algorithms or employing manual detection effort, and also the indirect opportunity costs of advertising revenue due to subsequent losses in user engagement [12, 13]. Therefore, platforms such as Facebook and Twitter face no compelling incentives to prevent the spread of disinformation. Thus, relying on platforms to police themselves will not work [14]. The only reason for a profit-motivated platform to control the spread of disinformation is to avoid penalties imposed either by users or a public regulator.

Techniques for mitigating disinformation must leverage tools in artificial intelligence (AI), which further complicates the issue of misaligned incentives. The sheer scale at which users generate and share content on social platforms mean that any form of content moderation must rely, to some degree, on the automation afforded by AI in order to handle the vast volume of data. This aspect is different from traditional publishing, television, and print media where humans are involved in the editorial feedback loop before any content is allowed to be published. The problem of assigning liability for content, thus, is also much simpler in traditional media. With user-generated content on social platforms, however, the quality and accessibility of data determine whether AI will be effective at moderating content. But because only platforms have full, real-time access to their data, the problem is thus to motivate their use of AI to proactively mitigate disinformation — in spite of their self-interest in not doing so [14] — without having the same expertise or access to data.

The *principal-agent framework* of microeconomics models the interactions between an *agent*, who can influence the probability of an outcome by incurring costly effort, and a *principal*, who has preferences over the outcome. We model the domain of disinformation prevention through this lens, with a platform as an agent who, via technological expertise and access to data, has the ability but not the incentive to undertake costly precautions against the spread of disinformation, and a regulator as the principal who seeks to balance the cost of precautions against the harm caused

by disinformation.

We begin by reviewing related work on techniques for mitigating disinformation and provide a background on the principal-agent model applied to regulation in Chapter 2. We then lay out our modeling assumptions in Chapter 3. We make three key assumptions. First, any attempt to regulate the actions of a platform before harm occurs (using a so-called negligence standard) must specify what level of effort for mitigating disinformation is adequate. Even if this specification is left implicit, we can model its effect as being a *public model*, operated by the regulator, requiring some level of effort. Second, any given public model will, in practice, require less effort over time as disinformation authors can learn to circumvent it; the platform is thus able to get away with expending less effort policing disinformation in order to save on costs, since the data and expertise needed to continuously re-train a model of content harmfulness is possessed by the platforms but not the regulators. Third, the public standard for content that ought to be prohibited on the platform will increase after a harmful event.

We formalize these assumptions as a Markov decision process (MDP) in Chapter 4, and use the model to derive our main results. We show that that no level of fines based on a static public model can induce optimal effort. However, in the presence of a public model that reacts to a harmful event by increasing the required level of effort, the platform's individually-optimal effort may exceed that currently required by the public model. In particular, the platform may be incentivized to continue exerting effort at a specific threshold when the public model becomes less stringent over time. However, perhaps counterintuitively, this effort threshold will fall short of the socially-optimal level unless the public model sufficiently *overreacts* as a response to any harmful event, by requiring a level of effort that is greater than socially optimal.

Finally, to further demonstrate the complexity of this incentive problem, we show that even under a simpler, more stylized setting — where the regulator has the same technical ability as the platform and the costs of harm from disinformation are known

— absent knowledge of the platform's costs of effort, there is no specification of the public model's required effort that will always incentivize the socially-optimal level of effort. We thus conclude that the design of mechanisms that may elicit the costs of foregone engagement incurred by the platform in policing disinformation is one of the promising directions for future work. Chapter 5 covers a detailed discussion of our main takeaways and additional prescriptions.

Chapter 2

Background and Related Work

In this chapter, we provide a brief overview of the microeconomic concepts that will aid in understanding interdisciplinary work of this kind. We start by situating this thesis within the disinformation prevention domain through a survey of existing techniques for combating false news online.

2.1 Fighting Falsity Online

Detecting disinformation. A popular approach towards limiting online disinformation is to develop tools or frameworks that are effective in detecting associated content. This process aims to identify disinformation in its initial stages so that mitigating efforts thereafter may restrict or eliminate exposure to users of social media. Zhou and Zafarani [15] survey some techniques that make false news detection efficient and explainable. These techniques are categorized into four areas: knowledge-based methods that involve fact-checking, style-based methods that focus on studying linguistic features of false content, propagation-based methods that analyze how such content spreads in the social network, and source-base methods that investigate the credibility of sources that generate false news. The goal of studying these and other characteristic features of the false news ecosystem, such as those surveyed by Kumar and Shah [1], is to develop algorithms and tools for early detection.

Knowledge-based methods mainly involve fact-checking, which in turn can be either

manual or automatic. Fact-checking is the process of extracting claims made in a given piece of content that is to be verified and checking these against known facts [15]. Manual fact-checking can either be crowd-sourced from users on social platforms, similar to Facebook or Twitter provisioning its users with the ability to report hoax content [16, 17]; or it can also be conducted via third-party websites such as *Snopes*[1], *PolitiFact*[2], or *FactCheck*[3] that employ domain experts dedicated to serving the public by debunking disinformation. Since manual fact-checking will not scale well with the volume of content generated on social media, work on developing automated fact-checking tools to assist human fact-checkers also exists [18]. The idea here is to develop a uniform structure to represent claims or statements in a given text so that it can be readily processed and compared to an existing set of facts. For instance, a simple representation of the form (Subject, Predicate, Object) may encode sentences like "Cristiano Ronaldo scored his 58th career hat-trick yesterday" as (CristianoRonaldo, goals, 3) [15].

While knowledge-based and style-based detection techniques focus on analyzing the textual content of disinformation — so that predictive classifiers might be trained to readily and effectively flag false news — propagation-based techniques study how such content disseminates amongst users in a given social network. Vosoughi *et al.* [19] have conducted an empirical study of tweets on Twitter to analyze the differences between the spread of true and false news stories. They codify the diffusion pattern of a news story in the form of a *cascade*, which is a simple tree-like representation where the root encodes the initial tweet and every subsequent level represents the how far (*tree depth*) and widespread (*tree breadth*) it spreads in the network. Such a representation helps compare false and true stories via precise measures: It is shown that false news stories travel faster, farther, and more widely than true news stories [19]. Such an analysis not only aids in investigating the causes and consequences of

[1]https://www.snopes.com
[2]https://www.politifact.com
[3]https://www.factcheck.org

disinformation proliferation, but it also helps formulate propagation-based false news detection as a classification problem: If users share a certain news story in a way that its diffusion cascade mimics that of previously spread false content, that story may potentially be flagged as requiring additional investigation.

Our work starts from the assumption that the platform has the ability to detect and limit the spread of objectionable content [20]. Our focus instead is on modeling the incentives faced by the platform to *not exercise this ability.*

Mitigating the effects of disinformation. Once effective technology for detecting disinformation content and diffusion networks is implemented, the next step is to mitigate or limit the impact such content may have on users. A straightforward approach is to simply remove associated content from the platform entirely; another is to demote or down-rank content so that it is less likely to be served on users' feeds. These approaches rely on platforms to undertake action to reduce the spread of disinformation since the recommendation algorithms serving content to users are proprietary. However, there are also studies conducted by third-party researchers offering other solutions for mitigating the effects of disinformation once it has entered the social network.

One such technique draws from concepts in human cognitive psychology to study deception cues that influence users' decision-making process related to sharing content in the social network [21]. The authors propose the implementation of plugins on social platforms that characterize content based on its acceptability (retweets in Twitter, for instance), source credibility (number of unique initial shares, for instance), message coherence, and message consistency, thereby guiding users about the veracity of content they encounter. The goal here is to provide users with informative cues so that they are less likely to share disinformation.

Another, more proactive intervention is the "Facts Before Rumors" campaign [22], where the focus is to preempt the kinds of rumours that are likely to spread on a

social network — based on user locations and localized news content, for example — and counteract these in advance by employing certain users to spread truthful news. This proactive methodology exploits the content diffusion network of a given social platform and set of users; and it seeks to expose enough users to truthful information so that they are less likely to believe in and forward false rumors.

Other interventions focus on curing the effects of disinformation instead of preventing it initially. For example, the "Correct the Record" initiative proposes a visual correction that may be sent to users exposed to false content on Facebook [23]. An advantage of this method is that it does not rely on predicting whether each piece of content contains disinformation as it is served to users, which is difficult due to problems of scale; instead, it allows platforms to be more reactive and debunk stories only after these are verified by experts as being false rumors. Moreover, the visual corrections proposed in [23] may seamlessly integrate with Facebook's existing user interface so as to provide the least possible friction to users exposed to disinformation.

Again, this thesis work focuses less on advocating a particular mitigation technique; rather, our goal is to determine the conditions under which platforms can be induced to actively implement any such technique to prevent harm from the spread of disinformation.

2.2 Hidden-Action Principal-Agent Model

Many economic interactions involve two parties, a principal and an agent, where the agent's choice of action imposes some form of (negative or positive) externality on the principal. In most realistic scenarios, the principal cannot directly monitor or observe the agent's action, but instead only observes a stochastic outcome resulting from it. For example, in the interaction between a property insurer (principal) and a property owner (agent), if the insurer bears the costs of any damages to the property, the owner might not be incentivized to maintain it and might engage in risky behaviors (e.g. leave the kitchen unattended while cooking). This situation exemplifies the problem of

moral hazard, which is an important feature of the principal-agent interaction because it precludes straightforward incentive schemes. Many employment settings also share this characteristic. For example, the CEO (principal) of a small startup company — whose income is directly related to the company's growth and product sales — would want their employees (agent) to undertake effort that profits the company (e.g. a UI/UX developer improving the company's website leading to increased traffic and sales). But if the employees are simply compensated at a fixed hourly rate, they might not be incentivized to put in their best effort to benefit the company.

Naturally, it will be in the principal's interests to influence the agent's choice of action. The principal may therefore be invested in drafting a contract for such influence in order to guard against the problem of moral hazard [24]. The need for a contract arises due to information asymmetry between the two parties — that is, the agent has more information or expertise about their actions than the principal. For property insurance, the hidden information is the agent's act of not maintaining the property and engaging in some risky behavior; for the startup company example, the hidden information is the UI/UX developer's expertise in developing clean, functional websites; whereas for our setting, AI is the source of asymmetric information: Only platforms possess the expertise, models and data to promptly flag and mitigate disinformation.

2.2.1 Contract Theory Meets Computer Science

The principal-agent model is central to *contract theory*, which is an important field in microeconomics. This area has recently gained traction in the algorithmic game theory community, primarily through works such as [24–26], where the aim is to concisely represent principal-agent settings and computationally characterize the design of *optimal* contracts[4] permitted by such settings. For example, Dütting *et al.* [24] contrast optimal contracts with their simple, *linear* counterparts for the classic principal-agent

[4]An optimal contract is one that maximizes the principal's expected reward assuming that the agent best responds to the contract [26].

situation [27] — that is, where the principal devises an outcome-dependent payment scheme to induce the agent to undertake some costly action — and show that while an optimal contract is straightforward to compute via linear programming, it is complex and unintuitive in practice. The authors explain the prevalence of simple contracts via a novel notion about their robustness, providing also the worst-case approximation guarantees of these contracts.

Our work is similar to these studies in that we consider the optimal design problem of maximizing the utility of the principal, who in our setting is a social welfare-maximizing regulator. Yet, instead of a computational complexity analysis for the design of contracts in classic principal-agent settings, we represent disinformation prevention as a principal-agent problem through our descriptive MDP model, which to our knowledge is a unique and first approach towards modeling the incentives faced by social platforms pertaining to the mitigation of false news and other toxic content. Unlike those cited works, the outcome space for our setting is simply the realization of a single harmful event due to the unmitigated spread of disinformation; our focus as such is specifically on the design of penalty contracts or schemes enforced by some regulatory agency in order to contain this stochastic externality, or harm from disinformation.

2.2.2 Regulating Stochastic Externalities

The hidden action principal-agent model can also be applied to the regulation of firms that generate stochastic externalities as a result of their operations. Examples include harmful accidents such as medical product failures, oil spills, nuclear waste leakages and other forms of pollution [28]. Moral hazard exists in these settings because firms (agent) might not be incentivized to take a costly precaution (unobservable action) to reduce accident risk, which is where a regulatory authority (principal) steps in to specify a penalty contract to guarantee some form of enforcement.

Cohen [29] explores optimal enforcement strategies for the regulation of firms that

stochastically pollute the environment in the form of oil spills. It is shown that under a *strict liability standard*, where a polluting firm is always penalized if an oil spill occurs regardless of its level of precautionary effort, the firm can be induced to exercise the socially-optimal or *first-best* level of effort. However, this requires that a specific firm can be identified as being responsible for a spill.

When a strict liability standard is impractical — for example because the perpetrator of harm cannot be reliably identified — a regulator might prefer to expend resources to monitor a firm's effort directly. In these situations, a *negligence standard* can be preferable, in which a firm is not held responsible for an accident if it can demonstrate that it took adequate precautions. Naturally, the quality of information available for regulatory monitoring is a consideration for enforcing such a standard [30].

Our domain shares many of the features of the oil spill prevention domain: There are stochastic externalities associated with the spread of certain kinds of content on social platforms (harm from disinformation), as there are with firms transporting oil (oil spills); the likelihood or severity of such harm may be reduced to some degree if platforms exercise responsible and proactive content moderation, but not completely eliminated as the harm is ultimately a direct outcome of individual actions — akin to a spill that occurs because of inclement weather and not due to the oil tanker being faulty. However, our domain is also importantly different from that of oil spill regulation, mainly because of the difficulty in specifying adequate precautions against disinformation and also due to the strategic nature of disinformation authors. The following section expands on these differences. In Chapter 3 that introduces our formal, descriptive model, we will elaborate on the similarities and highlight how these key differences prevent the application of standard enforcement strategies.

2.3 Why Online Disinformation is Different

Disinformation prevention via regulatory mechanisms has its own unique challenges. First, there are ongoing debates around assigning liability for content hosted by social platforms [14], particularly due to editorial control being different for the social media setting. As discussed previously, it is infeasible to implement human-in-the-loop feedback for every piece of real-time, user-generated content shared on online platforms, as this medium is unlike traditional forms of media; there exist as such not only the issue of scalabilty for any disinformation mitigation technology, but also the question about whether similar liability rules for harmful content should apply to social media as they would for traditional media.

Second, in order to handle the vast volume of content, AI must be utilized for the proactive and automated flagging of disinformation. This aspect complicates regulation because the data powering such AI is only accessible to the social platforms themselves. Moreover, the recommendation algorithms that filter and serve content to users are also proprietary. Therefore, unlike for the environmental regulation domain, mandating exact precautions against the spread of disinformation for social platforms is likely to be an involved process for any regulatory authority — especially in comparison to, for example, specifying precise conditions that render an oil tanker safe for the transport of oil, or promoting adequate technology that will reduce emissions causing air pollution.

Third, and also different from pollution regulation, there exists the issue of malicious actors responding strategically to any explicitly fixed rules or precautions against the spread of disinformation. Authors and purveyors of disinformation are constantly coming up with new, sophisticated methods to ensure that their fabricated stories disseminate online: Techniques include obfuscation strategies to hide disinformation propagating networks and the origins of propagandist content; and also changing the content itself via constructing new falsehoods, or targeting different groups [2, 31].

Any successful attempts to moderate such users or content at scale must therefore utilize all the technical expertise and data required to counteract efforts of these bad actors. Regulation becomes challenging because only social platforms have access to such resources and data, and they are not necessarily incentivized to undertake action at the expense of losses in user engagement [12, 14].

2.4 Mechanism Design

Another closely related body of work is the economic theory of mechanism design, where the goal is to design protocols or procedures that mediate interactions between strategic agents in order to achieve some desired objective. Naturally, the outcome is subject to the constraint that agents behave selfishly, in that they act according to their rational self-interests; and also that agents hold some private information, that is their *hidden types*. A mechanism seeks to attain the desired outcome by incentivizing agents to report their private types. Mechanism design theory contrasts with the standard principal-agent model with respect to where the information asymmetry exists: it is the agents' type information that is hidden from the mechanism designer; whereas, for the principal-agent model, the principal cannot directly observe an agent's action(s), which form(s) the source of asymmetric information.

Because this thesis is concerned with setting up a regulatory policy in order to achieve a desired social outcome — that is, the socially-optimal level of control of disinformation — mechanism design is a pertinent framework for our domain. As described previously, we consider the harm from the spread of disinformation as a negative externality inextricably linked with the usage of social media platforms, akin to pollution being a by-product of certain firms' production activities. Indeed, Baliga and Maskin [32] formally demonstrate that in the presence of *nonexcludable* externalities such as pollution, government intervention in the form of a mechanism is necessary to achieve Pareto-efficient outcomes. Yet, in applying classic findings of mechanism design literature to their pollution reduction model, the authors assume

that agents' pollution reduction efforts are verifiable by the government. This assumption is quite strong as it bypasses the problem of moral hazard entirely, which is central to the principal-agent setting and thus also a key feature of our domain.

Mechanism design, hence, does not directly apply to our scenario because the regulator cannot reliably observe a social platform's efforts to curb the spread of harmful content. We therefore utilize the principal-agent framework to model the regulation of disinformation. The goal for the regulator (principal) is to incentivize a platform (agent) to use its proprietary expertise and AI technology — which are not available to the regulator — to responsibly limit toxic and harmful content in order to control the harm from disinformation. The following chapter introduces our formal descriptive model.

Chapter 3

Modeling the Regulation of Disinformation

We have the following scenario: A regulator (principal) would like the platform (agent) to limit the amount of disinformation spread to control the likelihood of a stochastic and observable harmful event. The underlying assumption is that the unmitigated spread of disinformation on social platforms makes the occurrence of harm more likely.

We assume the platform possesses a proprietary classification model that accurately assigns for every a piece of content the probability of it causing harm [12, 20]. Thus, extremely violent, graphic, or objectionable content, which contains nudity, racism, child pornography, or any form of human/animal abuse, is tagged by the model with a very high harm probability value. Other, benign forms of content, such as cute photos of pets or birthday greetings, are assigned with a very low harm probability value.

Using this model, the platform can flag content exceeding some chosen harm probability threshold as being unacceptable and in violation of their community standards of acceptable postings. Thus, the platform's mitigation effort constitutes first detecting such content, and thereafter employing methods to either filter it entirely, downgrade it so that it appears on fewer user feeds, or label it with a warning invoking users' discretion. As discussed previously, the exact choice of technique is not

important for this analysis; any and all such methods effectively count as the platform exercising precautions against the spread of harmful content and, by extension, disinformation.

Let H be the binary random variable indicating whether harm occurs with density function $h(e) = \Pr[H \mid e] \in (0, 1]$ representing the probability that harm occurs if the platform exerts effort e. Similar to [29], we assume that although the platform is unable to control this externality directly, the platform can make it less likely for harm to occur by exercising more effort. In line with the standard economic model of unilateral accidents [28, 33, 34], we assume diminishing returns to effort — that is, effort reduces risk of harm at a decreasing rate: $h'(e) < 0$ and $h''(e) \geq 0$.[1]

The business model of social platforms is primarily based on online advertising generated when users engage with content by liking, clicking, and sharing [14, 35]. Thus, in addition to the direct costs of implementing content moderation, mitigating disinformation is costly due to the indirect costs of losing potential ad revenue. Let $c(e)$ denote the cost of exerting effort e. We assume effort is increasingly costly, that is $c'(e) > 0$ and $c''(e) > 0$, which is also standard under the unilateral accident model.

Given their behavioral advertising business model, platforms face no incentives to moderate attention-grabbing content, toxic or otherwise, especially because they do not directly incur the costs of any societal harm [12, 14]. Under this scenario of misaligned incentives, a social welfare-maximizing regulator aims to incentivize the platform to exercise adequate precautions against the spread of disinformation. Concretely, the regulator wishes to maximize the expected social welfare,

$$EW(e) = -h(e)D - c(e), \tag{3.1}$$

where D is the societal cost of harm (in dollars) due to disinformation, assumed to be constant here for simplicity.

The socially-optimal or first-best effort maximizing (3.1) is given by $e^* = \arg\max_e EW(e)$.

[1] Though the cited studies assume strict convexity of harm function, that is $h''(e) > 0$, our results are robust towards slightly relaxing this assumption.

At e^*, the sum of the total expected costs of harm, or $h(e^*)D$, and the platform's costs of exerting this effort, or $c(e^*)$, is minimized; thus, e^* by definition is the platform's precautionary effort at which the cost of any additional effort is balanced by the expected cost of damages due to harm.

We discuss possible regulatory schemes by which the platform is induced to exert effort e^*, and further expand on domain specific features for our descriptive model.

3.1 Strict Liability

Under the strict liability standard, the platform is held completely liable for any harmful event, irrespective of its precautionary effort. To incentivize the first-best level of effort e^*, the strict liability fines T must equal D, the societal cost of harm [29]; thus, the platform's expected utility is given by,

$$EU(e) = -c(e) - h(e)T, \tag{3.2}$$

which equals the expected social welfare equation (3.1).

A regulator might pick this enforcement standard because it does not require expending resources to monitor the platform's effort, which is only imperfectly observable because of the difficulty in identifying the exact mechanics of the platform's proprietary algorithms. However, strict liability is impractical for a few reasons. Most importantly, the direct causal links between any harmful event and the platform are sufficiently loose for this standard not to work, since the perpetrators are ultimately individuals; the platform can claim plausible deniability, or point to efforts at prohibiting dangerous content after the harm has already occurred, akin to when Facebook and Twitter banned groups like "QAnon" or "Proud Boys" after the insurrection on Capitol Hill [36]. Furthermore, it is also difficult to estimate D a priori as the harm could manifest in different forms.

3.2 Negligence

Under this standard, a regulator must specify a duty of care that the platform must follow in order to avoid liability for any harm. Monitoring the platform's effort is thus necessary to determine liability.

Although monitoring is imperfect, the platform's content moderation efforts are not completely unobservable: there exists a crude public notion about the kinds of content that ought to be limited on social platforms. From an incentive standpoint, a negligence standard already exists in the sense that there is not a lot of nudity or child pornography, or content with explicit death threats, vile or racist remarks on most social platforms — platforms like Facebook and Twitter expend ample resources to enforce their community standards via active content moderation [20, 37]. Presumably, platforms do not want public outrage, or to be charged with trafficking or any other forms of liability for such content, which if not controlled would be reported extensively in popular press.

We model this descriptive situation with the presence of an explicit public model, operated by a regulator, that fixes a required level of precautionary effort for mitigating disinformation. In reality, there is no concept of an explicit public model specifying effort, but rather an implicit public notion about the types of content that ought to be moderated by the platform. Nonetheless, regardless of what the public standards are at any given moment, these standards imply a certain level of precautionary effort, which we encode with an explicit public model to simplify our formal analysis.

3.3 Performative Prediction of Disinformation

When predictions about the actions of an agent influence outcomes for that agent, there is a risk that the predictive model will cease to be accurate. For example, a certain keyword that is extremely predictive of a message being spam may cease to

be predictive once we filter based on it, as spammers will now have an incentive to stop using that keyword. Predictions that exhibit this problem are *performative*[2] — the prediction influences the outcome [38]. Classifying disinformation is performative in this sense because bad actors can learn to bypass any detection model with new forms of disinformation [2].

We assume the platform has sufficient technical resources and the data to retrain its proprietary model in order to counterbalance performativity; that is, the platform is able to successfully classify future modifications of disinformation via predicting true harm probabilities of associated content. The same is not true for the public model as the regulator does not possess the same expertise or access to data. The regulator in theory could utilize open-source, state-of-the-art disinformation detection learning models to effectively flag false content as not satisfying the public standard [39, 40]. Yet, to the extent that platform data is not completely accessible to the public [41], these open-source models will be susceptible to performative prediction of new, evolved forms of disinformation unless retrained with the same, easily accessible data that is available to the platform.

Consequently, because it is publicly accessible, the public model weakens over time due to performativity as disinformation authors strategically learn to circumvent it. We encode this feature effectively as a gradual *downward drift* or decrease in the public model's required effort if harm does not occur. However, if harm occurs, we see a public backlash in that the public's tolerance of content linked to the harmful event gets lower ex post. This is akin to when Facebook and Twitter began suspending accounts, content, and hashtags linked to the Capitol Hill riots [42]. We encode this backlash as an effective increase in the public model's required effort as a response to a harmful event.

[2]Note that we use the terms performative and performativity in a specific, strictly technical sense that differs from their colloquial usage.

Chapter 4

Formal Model

We formalize our model as a MDP incorporating descriptive features of our domain as described in the previous section and defined by (S, A, P_e, R_e) where S is the discrete state space of the current effort e_c required by the public model, A is the continuous set of actions representing the platform's choice of effort e, $P_e(e_c, e'_c) = \Pr[s_{t+1} = e'_c \mid s_t = e_c, a_t = e]$ is the transition probability to state e'_c by exerting effort e in state e_c, and $R_e = -c(e)$ is the immediate reward of exerting effort e, which is simply the cost of effort e. Consistent with MDP literature [43], we use $\pi : S \to A$ to denote an arbitrary, deterministic policy specifying the platform's choice of effort $e \in A$ for all $e_c \in S$. The *state value function* $v_\pi(e_c) = R_e + \gamma \mathbb{E}[v_\pi(s_{t+1})]$ is the expected discounted value of following policy π from state e_c; the *state-action value function* $q_\pi(e_c, e) = R_e + \mathbb{E}[v_\pi(s_{t+1}) | A_t = e]$ is the expected discounted value of choosing effort e in state e_c, and then following policy π thereafter.

4.1 Optimal Effort Under a Static Public Model

In the first analysis, we assume a static standard model with no downward drift of the public model's required effort level and no backlash if harm occurs; that is, the effort required by the public model remains fixed at e_c. Under this negligence standard, the platform is only subject to *ex ante* regulation via regulatory audits, and not penalized *ex post* if harm occurs.

Let $r \in [0,1]$ be the probability that the regulator conducts an audit of the platform's effort and let $P_f(e \mid e_c) \in [0,1]$ be the probability that the platform fails its audit if it exerts effort e, given the current required effort e_c. If the platform fails the audit, it is liable for fines F. Thus, assuming risk-neutrality, the platform's expected utility under ex ante regulation is,

$$EU(e \mid e_c) = -c(e) - rP_f(e \mid e_c)F. \tag{4.1}$$

Definition 1 *The adequate level of effort e is the point beyond which the probability of failing the audit $P_f(e \mid e_c) = 0$, where e_c is the public model's prescribed effort.*

Note that effort e_c is considered adequate because it is specified by the regulator. Thus, by definition, $P_f(e \mid e_c) = 0$ for all $e \geq e_c$: the platform never fails its audit by at least following the public model's prescribed effort (full compliance). Now note that for any given e_c and fine structure F, there exists an individually-optimal level of effort that maximizes (4.1). By inspection, this individually-optimal level will never exceed e_c, irrespective of how large the size of fines F is.

Proposition 1 *Given a fixed adequate effort level e', there exists no fine scheme F that can incentivize the platform to exert more effort than e'.*

Proof. By contradiction. Suppose the platform prefers to exert effort $e > e'$. Thus, the following must hold:

$$\begin{aligned} & EU(e|e_c) > EU(e'|e_c) \\ \iff\ & -c(e) - rP_f(e|e_c)F > -c(e') - rP_f(e'|e_c)F \end{aligned}$$

By definition, $P_f(e'|e_c) = 0$ and therefore $P_f(e|e_c) = 0$. Thus,

$$\begin{aligned} & -c(e) - rP_f(e|e_c)F > -c(e') - rP_f(e'|e_c)F \\ \iff\ & -c(e) > -c(e') \\ \iff\ & c(e') > c(e) \end{aligned}$$

which does not hold for $e > e'$ because by assumption $c'(e) > 0$ for all e (contradiction). ■

This result trivially follows from the specified conditions: for any two adequate effort levels, the platform will pick lower effort because that will maximize (4.1). Thus, with the static public model, no amount of fines solely based on ex ante regulation, no matter how large, can induce the platform to exercise more effort than the public model's specified e_c. Only if $e_c = e^*$, and if the regulator can guarantee full compliance with the public model, can this scheme incentivize socially-optimal effort.

4.2 Optimal Effort Under an Adaptive Public Model

We now consider an adaptive MDP setup. The current state represents the required level of effort e_c; if no harm occurs, the required effort reduces over time due to performativity; and if harm does occur, then the required effort increases to e_h, representing public backlash.

Assumption 1 *Given a fixed fine structure F and the effort required by the public model e_c, the platform's individually-optimal effort level is at least e_c.*

This assumption is without loss of generality: we will label the states of the MDP according to the individually-optimal static effort required given F and e_c.[1]

Assumption 2 *At state e_c, the transition probability to the high effort state $e_h > e_c$ is simply $P_e(e_c, e_h) = h(e)$, the probability that harm occurs given the platform exercises effort e.*

Note that the harm probability and thus the transition to state e_h only depends on the platform's effort e, and not on the state e_c. This transition encodes the public backlash.

[1]The platform will thus never exert less than e_c effort in state e_c, but we will see that it will sometimes exert more.

Definition 2 *The next state with required effort lower than e_c is $\chi(e_c) = \sup\{s \in S \mid s < e_c\}$.*

Assumption 3 *If harm does not occur, we assume a weakening of the public model via a continuous downward drift of the public model's prescribed effort — that is, the effort either lowers to $\chi(e_c)$ with drift probability $g(e_c)$, or stays fixed at e_c with probability $1 - g(e_c)$.*

Note that the drift probability to state $\chi(e_c)$ only depends on the current state e_c, and not the platform's effort e, conditional on the harm's not occurring. The decrease in effort encodes performativity.

Lemma 1 *Fix a state e_c representing the current effort required by the public model, and an arbitrary policy π, and let $e_h > e_c$ be the effort that the public model will require if harm occurs. For all $e_2 > e_1 \geq e_c$,*

$$q_\pi(e_c, e_2) > q_\pi(e_c, e_1) \iff d(\pi, e_c) - v_\pi(e_h) > \frac{c(e_2) - c(e_1)}{\gamma(h(e_1) - h(e_2))}, \tag{4.2}$$

where $d(\pi, e_c) = g(e_c)v_\pi(\chi(e_c)) + (1 - g(e_c))v_\pi(e_c)$.

Proof. At e_c, the state-action value function for some effort e is given by:

$$\begin{aligned} q_\pi(e_c, e) &= \mathbb{E}_\pi[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = e_c, A_t = e] \\ &= \sum_{e'_c} P(e'_c \mid s = e_c, a = e)[-c(e) + \gamma v_\pi(e'_c)] \\ &= -c(e) + \gamma[h(e)v_\pi(e_h) + (1 - h(e))(g(e_c)v_\pi(\chi(e_c)) + (1 - g(e_c))v_\pi(e_c))] \end{aligned}$$

By substituting in $d(\pi, e_c) = g(e_c)v_\pi(\chi(e_c)) + (1 - g(e_c))v_\pi(e_c)$ we have:

$$q_\pi(e_c, e) = -c(e) + \gamma[h(e)v_\pi(e_h) + (1 - h(e))d(\pi, e_c)]. \tag{4.3}$$

Thus, for $q_\pi(e_c, e_2) > q_\pi(e_c, e_1)$, we have:

$$\begin{aligned}
& -c(e_2) + \gamma[h(e_2)v_\pi(e_h) + (1 - h(e_2))d(\pi, e_c)] > \\
& -c(e_1) + \gamma[h(e_1)v_\pi(e_h) + (1 - h(e_1))d(\pi, e_c)] \\
\iff \quad & -c(e_2) + \gamma[h(e_2)v_\pi(e_h) + d(\pi, e_c) - h(e_2)d(\pi, e_c)] > \\
& -c(e_1) + \gamma[h(e_1)v_\pi(e_h) + d(\pi, e_c) - h(e_1)d(\pi, e_c)] \\
\iff \quad & -c(e_2) + \gamma h(e_2)v_\pi(e_h) + \gamma d(\pi, e_c) - \gamma h(e_2)d(\pi, e_c) > \\
& -c(e_1) + \gamma h(e_1)v_\pi(e_h) + \gamma d(\pi, e_c) - \gamma h(e_1)d(\pi, e_c) \\
\iff \quad & -c(e_2) + \gamma h(e_2)v_\pi(e_h) - \gamma h(e_2)d(\pi, e_c) > \\
& -c(e_1) + \gamma h(e_1)v_\pi(e_h) - \gamma h(e_1)d(\pi, e_c) \\
\iff \quad & -c(e_2) - \gamma h(e_2)(d(\pi, e_c) - v_\pi(e_h)) > \\
& -c(e_1) - \gamma h(e_1)(d(\pi, e_c) - v_\pi(e_h)) \\
\iff & \gamma h(e_1)(d(\pi, e_c) - v_\pi(e_h)) - \gamma h(e_2)(d(\pi, e_c) - v_\pi(e_h)) > c(e_2) - c(e_1) \\
\iff & \gamma(h(e_1) - h(e_2))(d(\pi, e_c) - v_\pi(e_h)) > c(e_2) - c(e_1) \\
\iff & d(\pi, e_c) - v_\pi(e_h) > \frac{c(e_2) - c(e_1)}{\gamma(h(e_1) - h(e_2))}.
\end{aligned}$$

■

Given e_c, Lemma 1 specifies the condition under which the platform's picks one effort level over another from the continuous action set A, expressed via the state-action value function of the MDP.

Definition 3 *We call* π^τ *a* threshold strategy with threshold τ *if* $\pi^\tau(e_c) = \max\{\tau, e_c\}$ *for all* $e_c \in S$.

Threshold strategies form a class of policies that can induce more aggressive effort as specified by the condition in Lemma 1. The following results characterize important features of threshold strategies lending support to our main derivation of the platform's optimal policy in Theorem 1.

Lemma 2 *Given a threshold strategy π^τ, the state value function $v_{\pi^\tau}(e_c)$ is fixed for all $e_c \leq \tau$.*

Proof. The state value function for some arbitrary $e_c \leq \tau$ is given by,

$$v_{\pi^\tau}(e_c) = -c(\tau)+\gamma[h(\tau)v_{\pi^\tau}(e_h)+(1-h(\tau))[g(e_c)v_{\pi^\tau}(\chi(e_c))+(1-g(e_c))v_{\pi^\tau}(e_c)]]. \quad (4.4)$$

Note that the platform's policy specifying effort for all $e_c \leq \tau$ is fixed by definition; that is, $\pi^\tau(e_c) = \tau$ for all $e_c \leq \tau$. Thus, the transition to state e_h is also fixed because the transition probability $h(\tau)$ is fixed. And similarly, the probability that harm does not occur is also fixed at $(1 - h(\tau))$.

Let $e_0 = \min \mathcal{S}$. We prove inductively that $v_{\pi^\tau}(e_k) = v_{\pi^\tau}(e_0)$ for all $e_0 \leq e_k \leq \tau$. The base case ($v_{\pi^\tau}(e_0) = v_{\pi^\tau}(e_0)$) is immediate. For the inductive step, assume that $v_{\pi^\tau}(e_{k-1}) = v_{\pi^\tau}(e_0)$. Then

$$\begin{aligned} v_{\pi^\tau}(e_k) &= -c(\tau) + \gamma[h(\tau)v_{\pi^\tau}(e_h) + (1 - h(\tau))(1 - g(e_k))v_{\pi^\tau}(e_k) + (1 - h(\tau))g(e_k)v_{\pi^\tau}(e_{k-1})] \\ &= -c(\tau) + \gamma[h(\tau)v_{\pi^\tau}(e_h) + (1 - h(\tau))(1 - g(e_k))v_{\pi^\tau}(e_k) + (1 - h(\tau))g(e_k)v_{\pi^\tau}(e_0)]. \end{aligned}$$

Thus, $v_{\pi^\tau}(e_k) = g(e_k)V_1 + (1 - g(e_k))V_0$, where

$$\begin{aligned} V_1 &= -c(\tau) + \gamma[h(\tau)v_{\pi^\tau}(e_h) + (1 - h(\tau))v_{\pi^\tau}(e_{k-1})] \\ &= -c(\tau) + \gamma[h(\tau)v_{\pi^\tau}(e_h) + (1 - h(\tau))v_{\pi^\tau}(e_0)] \\ &= v_{\pi^\tau}(e_0) \end{aligned}$$

and

$$\begin{aligned} V_0 &= -c(\tau) + \gamma h(\tau)v_{\pi^\tau}(e_h) + \gamma(1 - h(\tau))v_{\pi^\tau}(e_k) \\ &= -c(\tau) + \gamma h(\tau)v_{\pi^\tau}(e_h) + \gamma(1 - h(\tau))[g(e_k)V_1 + (1 - g(e_k))V_0]. \end{aligned}$$

Note that the following is also true for V_1:

$$\begin{aligned}
V_1 &= -c(\tau) + \gamma[h(\tau)v_{\pi^\tau}(e_h) + (1-h(\tau))v_{\pi^\tau}(e_0)] \\
&= -c(\tau) + \gamma[h(\tau)v_{\pi^\tau}(e_h) + (1-h(\tau))g(e_k)v_{\pi^\tau}(e_0) + (1-h(\tau))(1-g(e_k))v_{\pi^\tau}(e_0)] \\
&= -c(\tau) + \gamma h(\tau)v_{\pi^\tau}(e_h) + \gamma(1-h(\tau))g(e_k)v_{\pi^\tau}(e_0) + \gamma(1-h(\tau))(1-g(e_k))v_{\pi^\tau}(e_0) \\
&= \sum_{j=0}^{\infty} \gamma^j (1-h(\tau))^j (1-g(e_k))^j [-c(\tau) + \gamma h(\tau)v_{\pi^\tau}(e_h) + \gamma(1-h(\tau))g(e_k)v_{\pi^\tau}(e_0)] \\
&= v_{\pi^\tau}(e_0)
\end{aligned}$$

for all $g(e_k) \in [0,1]$.

Thus, for V_0:

$$\begin{aligned}
V_0 &= -c(\tau) + \gamma h(\tau)v_{\pi^\tau}(e_h) + \gamma(1-h(\tau))[g(e_k)V_1 + (1-g(e_k))V_0] \\
&= -c(\tau) + \gamma h(\tau)v_{\pi^\tau}(e_h) + \gamma(1-h(\tau))[g(e_k)v_{\pi^\tau}(e_0) + (1-g(e_k))V_0] \\
&= -c(\tau) + \gamma h(\tau)v_{\pi^\tau}(e_h) + \gamma(1-h(\tau))g(e_k)v_{\pi^\tau}(e_0) + \gamma(1-h(\tau))(1-g(e_k))V_0 \\
&= \sum_{j=0}^{\infty} \gamma^j (1-h(\tau))^j (1-g(e_k))^j [-c(\tau) + \gamma h(\tau)v_{\pi^\tau}(e_h) + \gamma(1-h(\tau))g(e_k)v_{\pi^\tau}(e_0)] \\
&= V_1 \\
&= v_{\pi^\tau}(e_0).
\end{aligned}$$

But then

$$\begin{aligned}
v_{\pi^\tau}(e_k) &= g(e_k)V_1 + (1-g(e_k))V_0 \\
&= g(e_k)v_{\pi^\tau}(e_0) + (1-g(e_k))v_{\pi^\tau}(e_0) \\
&= v_{\pi^\tau}(e_0)
\end{aligned}$$

for all $g(e_k) \in [0,1]$, and we are done. ■

Lemma 2 fixes the reward of exerting effort at a specific threshold, thereby enabling a straightforward characterization and analysis of the platform's policy — amid all the possible drifting states of the public model — by means of a stable level of effort τ.

Proposition 2 *For all threshold strategies π^τ, we have that $v_{\pi^\tau}(e_h) \leq v_{\pi^\tau}(e_c)$ holds for all $e_c \in S$.*

Proof. For ease of notation, let $S = \{e^0, e^1, \ldots, e^h\}$ denote the set of all states with $e^0 < \cdots < e^h$, and let $\pi = \pi^\tau$ with $\tau = 0$. Note that this specification is w.l.o.g.; for $\tau > 0$, we will consider a subset of S such that the first state of this subset $e^0 = \sup\{e \in S \mid e \leq \tau\}$, since from Lemma 2 we know that the state value function for all $e \leq \tau$ is fixed.

Now we move on to the proof. Suppose the claim is false. Then $\{e \mid v_\pi(e) < v_\pi(e^h)\} \neq \varnothing$. Let $e^z = \min\{e \mid v_\pi(e) < v_\pi(e^h)\}$ and $d(e^j) = (1 - g(e^j))v_\pi(e^j) + g(e^j)v_\pi(e^{j-1})$ for all $0 \leq j \leq h$.

First, observe that

$$\begin{aligned} d(e^z) &= (1 - g(e^z))v_\pi(e^z) + g(e^z)v_\pi(e^{z-1}) \\ &\geq (1 - g(e^z))v_\pi(e^z) + g(e^z)v_\pi(e^z) \\ &= v_\pi(e^z), \end{aligned}$$

where the inequality follows from combining the assumptions $v_\pi(e^h) > v_\pi(e^z)$ with $v_\pi(e^{z-1}) \geq v_\pi(e^h)$, both from the definition of e^z. Note that if $e^z = e^0$, then the same result holds, since $g(e^0) = 0$.

It then follows that

$$\begin{aligned} v_\pi(e^z) &= -c(e^z) + \gamma[h(e^z)v_\pi(e^h) + (1 - h(e^z))d(e^z)] \\ &\geq -c(e^z) + \gamma[h(e^z)v_\pi(e^h) + (1 - h(e^z))v_\pi(e^z)] \\ &> -c(e^z) + \gamma[h(e^z)v_\pi(e^z) + (1 - h(e^z))v_\pi(e^z)] \\ &= -c(e^z) + \gamma v_\pi(e^z) \\ &\geq \sum_{j=0}^{\infty} \gamma^j(-c(e^z)). \end{aligned}$$

We now show inductively that $v_\pi(e^k) \geq v_\pi(e^h)$ for all $z \leq k < h$.

The base case is e^{h-1}. Suppose the contrary that $v_\pi(e^{h-1}) < v_\pi(e^h)$. Then we have

$$\begin{aligned} d(e^h) &= (1 - g(e^h))v_\pi(e^h) + g(e^h)v_\pi(e^{h-1}) \\ &\leq v_\pi(e^h), \end{aligned}$$

because $0 \leq g(e^h) \leq 1$, which gives

$$\begin{aligned} v_\pi(e^h) &= -c(e^h) + \gamma[h(e^h)v_\pi(e^h) + (1 - h(e^h))d(e^h)] \\ &\leq -c(e^h) + \gamma[h(e^h)v_\pi(e^h) + (1 - h(e^h))v_\pi(e^h)] \\ &= -c(e^h) + \gamma v_\pi(e^h) \\ &\leq \sum_{j=1}^{\infty} \gamma^j(-c(e^h)) \\ &< \sum_{j=1}^{\infty} \gamma^j(-c(e^z)) \\ &< v_\pi(e^z), \end{aligned}$$

contradicting the definition of e^z.

For the inductive step, assume that $v_\pi(e^k) \geq v_\pi(e^h)$, for some $z < k < h$. Then we show that $v_\pi(e^{k-1}) \geq v_\pi(e^h)$. Assume not; then similarly we have

$$\begin{aligned} d(e^k) &= (1 - g(e^k))v_\pi(e^k) + g(e^k)v_\pi(e^{k-1}) \\ &\leq (1 - g(e^k))v_\pi(e^k) + g(e^k)v_\pi(e^h) \\ &\leq (1 - g(e^k))v_\pi(e^k) + g(e^k)v_\pi(e^k) \\ &= v_\pi(e^k) \end{aligned}$$

and thus

$$\begin{aligned}
v_\pi(e^k) &= -c(e^k) + \gamma[h(e^k)v_\pi(e^h) + (1 - h(e^k))d(e^k)] \\
&\leq -c(e^k) + \gamma[h(e^k)v_\pi(e^h) + (1 - h(e^k))v_\pi(e^k)] \\
&\leq -c(e^k) + \gamma[h(e^k)v_\pi(e^k) + (1 - h(e^k))v_\pi(e^k)] \\
&= -c(e^k) + \gamma v_\pi(e^k) \\
&\leq \sum_{j=1}^{\infty} \gamma^j(-c(e^k)) \\
&< \sum_{j=1}^{\infty} \gamma^j(-c(e^z)) \\
&< v_\pi(e^z) \\
&< v_\pi(e^h) \\
&\leq v_\pi(e^k),
\end{aligned}$$

again yielding a contradiction.

Therefore, $v_\pi(e^k) \geq v_\pi(e^h)$ is true for all $z \leq k < h$, which in particular implies that the initial claim $\{e \mid v_\pi(e) < v_\pi(e^h)\} \neq \varnothing$ must be false, thus completing the proof. ■

Proposition 2 establishes e_h as the worst state for the platform following a threshold strategy. Intuitively, because it encodes the public backlash, e_h by definition is the highest effort the public model will require and thus it must also yield the lowest expected discounted reward for the platform. Crucially, this guarantee of the lowest reward in state e_h acts as the incentivizing mechanism for the platform to exert more effort than explicitly required by the public model.

Given these MDP dynamics of performativity and public backlash as a response to harm, we characterize the platform's individually-optimal effort policy at any state e_c. The following **existing results** support our main result in Theorem 1.

Lemma 3 (Boyd and Vandenberghe [44]) *Suppose f is a differentiable function*

of one variable in dom(f). Then f is convex if and only if

$$f(y) - f(x) \geq f'(x)(y - x)$$

holds for all $x, y \in$ dom(f). And analogously for strict convexity,

$$f(y) - f(x) > f'(x)(y - x) \tag{4.5}$$

for all $x \neq y$.

Lemma 4 (Sutton and Barto [43]) *Given a pair of deterministic policies π and π' such that for all states $s \in S$*

$$q_\pi(s, \pi'(s)) \geq v_\pi(s),$$

then $v_{\pi'}(s) \geq v_\pi(s)$.

We now specify the platform's individually-optimal policy under the adaptive MDP setup: the public model's prescribed effort e_c increases to the high effort state e_h if harm occurs; and e_c decreases over time conditional on harm not occurring. The following theorem demonstrates that the optimal policy for the platform under these dynamics is to follow a threshold strategy.

Theorem 1 *The optimal strategy π^* for the platform is a threshold strategy $\pi^* = \pi^{\hat{e}}$, with threshold*

$$\hat{e} = \sup \left\{ e \in [0, e_h] \,\middle|\, q_{\pi^e}(s^{-1}(e), e) - q_{\pi^e}(e_h, \pi^e(e_h)) \geq -\frac{c'(e)}{\gamma h'(e)} \right\}, \tag{4.6}$$

where $s^{-1}(e) = \sup\{e_c \in S \mid e_c \leq e\}$.

Proof. By contradiction. Suppose $\pi^{\hat{e}}$ is suboptimal. Then by the process of policy improvement, there must exist a state e_c where some effort $e \neq \pi^{\hat{e}}(e_c)$ guarantees a higher expected reward than $\pi^{\hat{e}}(e_c)$. Thus, we apply the policy improvement theorem (Lemma 4) to find any such e_c where $q_{\pi^{\hat{e}}}(e_c, e) > v_{\pi^{\hat{e}}}(e_c)$ holds, which would imply that a greedy deviation from $\pi^{\hat{e}}$ exists as the better policy.

Case 1 ($\forall e_c$): Less aggressive effort than e_c The first deviation from $\pi^{\hat{e}}$ at any e_c might be to exert less aggressive effort $e < e_c$. Suppose that less aggressive effort e guarantees a higher expected reward than the required effort e_c. However, we know that lower effort than e_c does not guarantee a higher expected reward for all e_c because e_c by definition is the platform's individually-optimal level of effort. Thus, we have a contradiction and this deviation does not work.

Case 2 ($e_c \leq \hat{e}$): Less aggressive effort than $\hat{e}$ Suppose that the platform prefers to exert less aggressive effort e_1 such that $e_c \leq e_1 < \hat{e}$. Then $q_{\pi^{\hat{e}}}(e_c, e_1) > q_{\pi^{\hat{e}}}(e_c, \hat{e})$ must be true.

Thus, $q_{\pi^{\hat{e}}}(e_c, \hat{e}) > q_{\pi^{\hat{e}}}(e_c, e_1)$ *must not* be true (contrapositive); or, by substituting in equation (4.2) from Lemma 1, the following must not hold:

$$d(\pi^{\hat{e}}, e_c) - v_{\pi^{\hat{e}}}(e_h) > \frac{c(\hat{e}) - c(e_1)}{\gamma(h(e_1) - h(\hat{e}))}. \tag{4.7}$$

From the definition in (4.6), note that because $\hat{e}$ is the supremum taken over a closed interval, it satisfies the following equation (*intermediate value theorem*):

$$q_{\pi^{\hat{e}}}(s^{-1}(\hat{e}), \hat{e}) - q_{\pi^{\hat{e}}}(e_h, e_h) = -\frac{c'(\hat{e})}{\gamma h'(\hat{e})}. \tag{4.8}$$

Now consider the L.H.S of (4.7) and of (4.8). Recall that $d(\pi^{\hat{e}}, e_c) = g(e_c)v_{\pi^{\hat{e}}}(\chi(e_c)) + (1-g(e_c))v_{\pi^{\hat{e}}}(e_c)$. Since $\pi^{\hat{e}}(e_c) = \hat{e}$ is fixed for all $e_c < \hat{e}$, the value functions $v_{\pi^{\hat{e}}}(\chi(e_c))$ and $v_{\pi^{\hat{e}}}(e_c)$ must be equal (Lemma 2). Thus, $d(\pi^{\hat{e}}, e_c) = v_{\pi^{\hat{e}}}(e_c)$ as $0 \leq g(e_c) \leq 1$. Furthermore, because $s^{-1}(\hat{e}) < \hat{e}$ by definition, $q_{\pi^{\hat{e}}}(s^{-1}(\hat{e}), \hat{e}) = v_{\pi^{\hat{e}}}(e_c)$ must be true. Moreover, $v_{\pi^{\hat{e}}}(e_h) = q_{\pi^{\hat{e}}}(e_h, e_h)$ as $e_h \geq \hat{e}$. Thus, the L.H.S of (4.7) and of (4.8) are equal, or

$$d(\pi^{\hat{e}}, e_c) - v_{\pi^{\hat{e}}}(e_h) = q_{\pi^{\hat{e}}}(s^{-1}(\hat{e}), \hat{e}) - q_{\pi^{\hat{e}}}(e_h, e_h). \tag{4.9}$$

Suppose that the following is true of the R.H.S of (4.7) and (4.8):

$$-\frac{c'(\hat{e})}{\gamma h'(\hat{e})} > \frac{c(\hat{e}) - c(e_1)}{\gamma(h(e_1) - h(\hat{e}))} \tag{4.10}$$

Thus,

$$\begin{aligned} & -\frac{c'(\hat{e})}{\gamma h'(\hat{e})} > \frac{c(\hat{e}) - c(e_1)}{\gamma(h(e_1) - h(\hat{e}))} \\ \iff & -\frac{c'(\hat{e})}{h'(\hat{e})} > \frac{c(\hat{e}) - c(e_1)}{h(e_1) - h(\hat{e})} \\ \iff & -\frac{c'(\hat{e})(e_1 - \hat{e})}{h'(\hat{e})(e_1 - \hat{e})} > -\frac{c(e_1) - c(\hat{e})}{h(e_1) - h(\hat{e})} \\ \iff & \frac{c(e_1) - c(\hat{e})}{h(e_1) - h(\hat{e})} > \frac{c'(\hat{e})(e_1 - \hat{e})}{h'(\hat{e})(e_1 - \hat{e})} \end{aligned}$$

Notice that the final inequality is always true: we know by assumption that c is strictly convex ($c''(e) > 0$) and so from equation (4.5) in Lemma 3 it follows that the numerator of the L.H.S must be strictly greater than the numerator of the R.H.S, that is $c(e_1) - c(\hat{e}) > c'(\hat{e})(e_1 - \hat{e})$; similarly, because h is convex ($h''(e) \geq 0$), the denominator of the L.H.S must be weakly greater than the denominator of the R.H.S, that is $h(e_1) - h(\hat{e}) \geq h'(\hat{e})(e_1 - \hat{e})$. Since $h'(e) < 0$ and $e_1 < \hat{e}$, it follows that the L.H.S fraction overall is strictly greater (*less negative*) than the R.H.S fraction (*more negative*).

Therefore, if (4.10) holds, then condition (4.7) must also hold because:

$$\begin{aligned} q_{\pi^{\hat{e}}}(s^{-1}(\hat{e}), \hat{e}) - q_{\pi^{\hat{e}}}(e_h, e_h) &= -\frac{c'(\hat{e})}{\gamma h'(\hat{e})} \\ \iff d(\pi^{\hat{e}}, e_c) - v_{\pi^{\hat{e}}}(e_h) &= -\frac{c'(\hat{e})}{\gamma h'(\hat{e})} \\ &> \frac{c(\hat{e}) - c(e_1)}{\gamma(h(e_1) - h(\hat{e}))}. \end{aligned}$$

If (4.7) holds, the contrapositive statement is false, and so the original statement must also be false; thus, the platform instead prefers to exactly exert effort $\hat{e}$, and no less, for all $e_c < \hat{e}$, a contradiction.

Case 3 ($e_c \leq \hat{e}$): More aggressive effort than $\hat{e}$ Suppose the platform prefers to exert excessive effort at some $e_2 > \hat{e}$. It follows that $q_{\pi^{\hat{e}}}(e_c, e_2) > q_{\pi^{\hat{e}}}(e_c, \hat{e})$ must hold, and so we have (Lemma 1):

$$d(\pi^{\hat{e}}, e_c) - v_{\pi^{\hat{e}}}(e_h) > \frac{c(e_2) - c(\hat{e})}{\gamma(h(\hat{e}) - h(e_2))}, \tag{4.11}$$

must also hold. Recall from (4.9) that,

$$d(\pi^{\hat{e}}, e_c) - v_{\pi^{\hat{e}}}(e_h) = q_{\pi^{\hat{e}}}(s^{-1}(\hat{e}), \hat{e}) - q_{\pi^{\hat{e}}}(e_h, e_h).$$

Thus,

$$\begin{aligned} d(\pi^{\hat{e}}, e_c) - v_{\pi^{\hat{e}}}(e_h) &> \frac{c(e_2) - c(\hat{e})}{\gamma(h(\hat{e}) - h(e_2))} \\ \iff q_{\pi^{\hat{e}}}(s^{-1}(\hat{e}), \hat{e}) - q_{\pi^{\hat{e}}}(e_h, e_h) &> \frac{c(e_2) - c(\hat{e})}{\gamma(h(\hat{e}) - h(e_2))} \end{aligned}$$

We know from (4.8) that,

$$q_{\pi^{\hat{e}}}(s^{-1}(\hat{e}), \hat{e}) - q_{\pi^{\hat{e}}}(e_h, e_h) = -\frac{c'(\hat{e})}{\gamma h'(\hat{e})}.$$

Thus, in order to guarantee that (4.11) holds, the following must be true:

$$\begin{aligned} -\frac{c'(\hat{e})}{\gamma h'(\hat{e})} &> \frac{c(e_2) - c(\hat{e})}{\gamma(h(\hat{e}) - h(e_2))} \\ \iff \quad -\frac{c'(\hat{e})}{h'(\hat{e})} &> \frac{c(e_2) - c(\hat{e})}{h(\hat{e}) - h(e_2)} \\ \iff -\frac{c'(\hat{e})(e_2 - \hat{e})}{h'(\hat{e})(e_2 - \hat{e})} &> -\frac{c(e_2) - c(\hat{e})}{h(e_2) - h(\hat{e})} \\ \iff \quad \frac{c(e_2) - c(\hat{e})}{h(e_2) - h(\hat{e})} &> \frac{c'(\hat{e})(e_2 - \hat{e})}{h'(\hat{e})(e_2 - \hat{e})} \end{aligned}$$

However, notice that this inequality *does not* hold for $e_2 > \hat{e}$: since c is strictly convex, we know from Lemma 3 that the numerator of the L.H.S is strictly greater than that of the R.H.S, that is $c(e_2) - c(\hat{e}) > c'(\hat{e})(e_2 - \hat{e})$; and similarly, because h is convex, the denominator of the L.H.S is weakly greater than that of the R.H.S, that is $h(e_2) - h(\hat{e}) \geq h'(\hat{e})(e_2 - \hat{e})$. Since $h'(e) < 0$ and $\hat{e} < e_2$, it follows that the L.H.S fraction overall must be strictly smaller (*more negative*) than the R.H.S fraction (*less negative*), that is,

$$\frac{c(e_2) - c(\hat{e})}{h(e_2) - h(\hat{e})} < \frac{c'(\hat{e})(e_2 - \hat{e})}{h'(\hat{e})(e_2 - \hat{e})}$$

must be true, a contradiction.

Case 4 ($e_c > \hat{e}$): More aggressive effort than e_c We prove an intermediate result to arrive at our contradiction for this case. We first show that $\hat{e}$ is the optimal effort threshold for all threshold strategies.

Let τ be the smallest $\tau > \hat{e}$ satisfying $q_{\pi^\tau}(e, \pi^\tau(e)) \geq q_{\pi^{\hat{e}}}(e, \pi^{\hat{e}}(e))$ for all $e \in S$. Let $s^{-1}(\tau) = e_1 < \tau$. Then following the definition of $\hat{e}$ in (4.6), we have

$$\begin{aligned} q_{\pi^\tau}(e_1, \tau) - v_{\pi^\tau}(e^h) &< \frac{-c'(\tau)}{\gamma h'(\tau)} \\ \iff q_{\pi^\tau}(e_1, \tau) - v_{\pi^\tau}(e^h) &< \frac{-c'(\sigma)}{\gamma h'(\sigma)} \end{aligned}$$

for $e_1 < \sigma < \tau$ and $\tau - \sigma$ sufficiently small. But since $d(\pi^\tau, e_1) = q_{\pi^\tau}(e_1, \tau)$ (Lemma 2), we have

$$\begin{aligned} d(\pi^\tau, e_1) - v_{\pi^\tau}(e^h) &< \frac{-c'(\sigma)}{\gamma h'(\sigma)} \\ &< \frac{c(\tau) - c(\sigma)}{\gamma(h(\sigma) - h(\tau))}, \end{aligned}$$

which implies by Lemma 1 that $q_{\pi^\tau}(e_1, \sigma) \geq q_{\pi^\tau}(e_1, \tau)$, and hence by the policy improvement theorem, $v_{\pi^\sigma}(e) \geq v_{\pi^\tau}(e)$ for all $e \in S$, contradicting the definition of τ. Hence there is no such threshold $\tau > \hat{e}$, and so $\hat{e}$ is the optimal threshold among all threshold strategies.

Now suppose the platform prefers to exert more aggressive effort at $e_2 > e_c$ for some $e_c > \hat{e}$. Thus, by Lemma 1, the following must hold:

$$\begin{aligned} d(\pi^{\hat{e}}, e_c) - v_{\pi^{\hat{e}}}(e_h) &> \frac{c(e_2) - c(e_c)}{\gamma(h(e_c) - h(e_2))} \\ \iff g(e_c) v_{\pi^{\hat{e}}}(\chi(e_c)) + (1 - g(e_c)) v_{\pi^{\hat{e}}}(e_c) - v_{\pi^{\hat{e}}}(e_h) &> \frac{c(e_2) - c(e_c)}{\gamma(h(e_c) - h(e_2))}. \end{aligned}$$

Thus, we have:

$$\begin{aligned} g(e_c)v_{\pi^{\hat{e}}}(\chi(e_c)) + (1-g(e_c))v_{\pi^{\hat{e}}}(e_c) - v_{\pi^{\hat{e}}}(e_h) &> \frac{c(e_2)-c(e_c)}{\gamma(h(e_c)-h(e_2))} \\ &> -\frac{c'(e_c)}{\gamma h'(e_c)} \\ &> -\frac{c'(\hat{e})}{\gamma h'(\hat{e})} \\ &= q_{\pi^{\hat{e}}}(s^{-1}(\hat{e}),\hat{e}) - q_{\pi^{\hat{e}}}(e_h,e_h) \\ &= v_{\pi^{\hat{e}}}(s^{-1}(\hat{e})) - v_{\pi^{\hat{e}}}(e_h). \end{aligned}$$

Thus,

$$\begin{aligned} & g(e_c)v_{\pi^{\hat{e}}}(\chi(e_c)) + (1-g(e_c))v_{\pi^{\hat{e}}}(e_c) - v_{\pi^{\hat{e}}}(e_h) > v_{\pi^{\hat{e}}}(s^{-1}(\hat{e})) - v_{\pi^{\hat{e}}}(e_h) \\ \iff \quad & g(e_c)v_{\pi^{\hat{e}}}(\chi(e_c)) + (1-g(e_c))v_{\pi^{\hat{e}}}(e_c) > v_{\pi^{\hat{e}}}(s^{-1}(\hat{e})), \end{aligned}$$

which implies that $v_{\pi^{\hat{e}}}(e_c) > v_{\pi^{\hat{e}}}(s^{-1}(\hat{e}))$ and/or $v_{\pi^{\hat{e}}}(\chi(e_c)) > v_{\pi^{\hat{e}}}(s^{-1}(\hat{e}))$. Thus, it follows that a new threshold strategy with threshold strictly greater than $\hat{e}$ will be preferable to $\hat{e}$, since exerting more aggressive effort e_2 in state e_c such that $e_2 > e_c > \hat{e}$ yields a better value. However, this implication contradicts our intermediate result because no threshold greater than $\hat{e}$ is optimal and we are done.

The process of policy improvement must give us a strictly better policy except when the original policy is already optimal [43]. Since there exists no greedy deviation $e \neq \pi^{\hat{e}}(e_c)$ such that $q(e_c,e) > q(e_c,\pi^{\hat{e}}(e_c))$ is true for any e_c, the proposed policy $\pi^{\hat{e}}$ must be optimal, thus completing the proof. ■

The result follows from the first-order condition of convexity [44] and the policy improvement theorem [43]. Intuitively, the theorem statement holds because past a certain level of effort, the gain to the platform of not exerting more effort is traded off against the increased probability of transitioning to the e_h state, which yields the lowest expected reward as shown in Proposition 2.

The primary takeaway from Theorem 1 is that the platform is incentivized to exert more aggressive effort at threshold $\hat{e}$, despite an over-time reduction of the public

model's prescribed effort e_c due to the performative prediction of disinformation. Thus, the platform's optimal effort level is stable at $\hat{e}$ for all states $e_c \leq \hat{e}$. The regulatory scheme that induces more aggressive effort is the ex post public backlash, that is when the required effort increases to e_h, which effectively poses as stricter future ex ante regulation as a response to a harmful event. This result is also important because with the correct choice of public backlash e_h, the platform can in theory be induced to exert the socially optimal level of effort e^*. We formalize this claim in the following proposition.

Proposition 3 *For any given socially optimal level of effort e^*, there exists a MDP consistent with our given conditions such that the optimal policy for the platform is a threshold strategy with threshold $\tau = e^*$.*

Proof. We know from Theorem 1 that under the specified conditions, the platform's optimal effort at any state e_c is a threshold strategy with threshold $\tau = \hat{e}$. In order to induce e^* as the optimal threshold, $\hat{e}$ must equal e^*; thus, from the defining constraint in (4.6), there must exist some e_h such that the following holds:

$$\begin{aligned} q_{\pi^{e^*}}(s^{-1}(e^*), e^*) - q_{\pi^{e^*}}(e_h, \pi^{e^*}(e_h)) &= -\frac{c'(e^*)}{\gamma h'(e^*)} \\ \iff v_{\pi^{e^*}}(e_0) - v_{\pi^{e^*}}(e_h) &= -\frac{c'(e^*)}{\gamma h'(e^*)}. \end{aligned} \tag{4.12}$$

where $e_0 = s^{-1}(e^*) \leq e^*$ (by definition).

Thus, we have

$$\begin{aligned} v_{\pi^{e^*}}(e_0) &= -c(e^*) + \gamma[h(e^*)v_{\pi^{e^*}}(e_h) + (1 - h(e^*))v_{\pi^{e^*}}(e_0)] \\ v_{\pi^{e^*}}(e_0) &= -c(e^*) + \gamma h(e^*)v_{\pi^{e^*}}(e_h) + \gamma(1 - h(e^*))v_{\pi^{e^*}}(e_0) \\ v_{\pi^{e^*}}(e_0) - \gamma(1 - h(e^*))v_{\pi^{e^*}}(e_0) &= -c(e^*) + \gamma h(e^*)v_{\pi^{e^*}}(e_h) \\ v_{\pi^{e^*}}(e_0)[1 - \gamma(1 - h(e^*))] &= -c(e^*) + \gamma h(e^*)v_{\pi^{e^*}}(e_h), \end{aligned}$$

and finally

$$v_{\pi^{e^*}}(e_0) = \frac{-c(e^*) + \gamma h(e^*)v_{\pi^{e^*}}(e_h)}{1 - \gamma(1 - h(e^*))}. \tag{4.13}$$

By substituting (4.13) in (4.12), we have

$$\begin{aligned}
v_{\pi^{e^*}}(e_0) - v_{\pi^{e^*}}(e_h) &= -\frac{c'(e^*)}{\gamma h'(e^*)} \\
\frac{-c(e^*) + \gamma h(e^*) v_{\pi^{e^*}}(e_h)}{1-\gamma(1-h(e^*))} - v_{\pi^{e^*}}(e_h) &= -\frac{c'(e^*)}{\gamma h'(e^*)} \\
-c(e^*) + \gamma h(e^*) v_{\pi^{e^*}}(e_h) - (1-\gamma(1-h(e^*)))v_{\pi^{e^*}}(e_h) &= -\frac{c'(e^*)(1-\gamma(1-h(e^*)))}{\gamma h'(e^*)} \\
-v_{\pi^{e^*}}(e_h)[-\gamma h(e^*) + 1 - \gamma(1-h(e^*))] &= c(e^*) - \frac{c'(e^*)(1-\gamma(1-h(e^*)))}{\gamma h'(e^*)} \\
-v_{\pi^{e^*}}(e_h)[-\gamma h(e^*) + 1 - \gamma + \gamma h(e^*))] &= c(e^*) - \frac{c'(e^*)(1-\gamma(1-h(e^*)))}{\gamma h'(e^*)} \\
-v_{\pi^{e^*}}(e_h)(1-\gamma) &= c(e^*) - \frac{c'(e^*)(1-\gamma(1-h(e^*)))}{\gamma h'(e^*)},
\end{aligned}$$

and finally

$$-v_{\pi^{e^*}}(e_h) = (\frac{1}{1-\gamma})(c(e^*) - \frac{c'(e^*)(1-\gamma(1-h(e^*)))}{\gamma h'(e^*)}). \tag{4.14}$$

We show by the *intermediate value theorem* (IVT) that (4.14) holds for some $e_h \in (e_{min}, e_{max})$ where e_{min} and e_{max} correspond to the lowest and highest possible levels of effort, respectively.

Let $G(e_h) = -v_{\pi^{e^*}}(e_h) - K$ where $K = (\frac{1}{1-\gamma})(c(e^*) - \frac{c'(e^*)(1-\gamma(1-h(e^*)))}{\gamma h'(e^*)})$. First, observe that $-v_{\pi^{e^*}}(e_h) \in (c(e_h), \frac{c(e_h)}{1-\gamma})$ by construction. Moreover, note that $K > 0$. Thus, we have the following at the lower bound of e_h:

$$\begin{aligned}
G(e_{min}) &= -v_{\pi^{e^*}}(e_{min}) - K \\
&< \frac{c(e_{min})}{1-\gamma} - (\frac{1}{1-\gamma})(c(e^*) - \frac{c'(e^*)(1-\gamma(1-h(e^*)))}{\gamma h'(e^*)}) \\
&= \frac{c(e_{min})}{1-\gamma} - \frac{c(e^*)}{1-\gamma} + (\frac{1}{1-\gamma})(\frac{c'(e^*)(1-\gamma(1-h(e^*)))}{\gamma h'(e^*)}) \\
&< 0,
\end{aligned}$$

for all $e^* \geq e_{min}$.

And we have the following at the upper bound of e_h:

$$\begin{aligned} G(e_{max}) &= -v_{\pi^{e^*}}(e_{max}) - K \\ &> c(e_{max}) - K \\ &> 0, \end{aligned}$$

which holds because the cost of exerting maximum possible effort, or $c(e_{max})$, is sufficiently large (by assumption).

Hence, because $G(e_{min}) < 0 < G(e_{max})$, it follows by the IVT that there exists some $e_h \in (e_{min}, e_{max})$ such that

$$\begin{aligned} & G(e_h) = 0 \\ \iff & -v_{\pi^{e^*}}(e_h) - K = 0 \\ \iff & -v_{\pi^{e^*}}(e_h) = K \\ \iff & -v_{\pi^{e^*}}(e_h) = (\frac{1}{1-\gamma})(c(e^*) - \frac{c'(e^*)(1-\gamma(1-h(e^*)))}{\gamma h'(e^*)}), \end{aligned}$$

and we are done. ■

The existence proof for e_h directly follows from the defining constraint of the platform's optimal policy in (4.6) and the continuity assumptions of the cost and harm functions. An interesting consequence of this result, however, is captured in the following proposition, where we effectively specify a strict lower bound on the public backlash as a necessary condition to induce the socially-optimal effort e^*.

Proposition 4 *The platform's optimal stable effort is guaranteed to be socially sub-optimal unless the public model overreacts by requiring effort $e_h > e^*$ if harm occurs.*

Proof. This result directly follows from the defining condition of the platform's stable effort $\hat{e}$ in (4.6). For $\hat{e}$ to equal e^*, the required effort e_h must be strictly greater than e^*. By definition, $\hat{e}$ is the supremum over the closed interval $[0, e_h]$ and so if $e_h < e^*$, then $\hat{e} < e^*$ is also true.

If $e_h = e^*$, then $\hat{e} < e^*$ is also true; the L.H.S of (4.6) equals zero for $e = e_h$, or

$$q_{\pi^{e_h}}(s^{-1}(e_h), e_h) - q_{\pi^{e_h}}(e_h, \pi^{e_h}(e_h)) = 0,$$

while the R.H.S is always positive, or

$$-\frac{c'(e_h)}{\gamma h'(e_h)} > 0,$$

since $c'(e) > 0$ and $h'(e) < 0$ for all $e \in [0, 1]$, and therefore the inequality is not satisfied. Thus, $e_h > e^*$ must be true in order for the platform's stable effort $\hat{e}$ to equal e^*. ■

Our ancillary result in Proposition 4 captures the counterintuitive nature of the penalty scheme according to our model: it is not sufficient to set the ex post required effort to the optimal effort e^*, assuming e^* were known; instead, to incentivize optimal effort, the public model must overreact and mandate suboptimal effort $e_h > e^*$ as a response to any harmful event.

4.3 Incentivizing Socially-Optimal Effort Under a Robust Public Model

Our descriptive model requires overreacting to harmful events in order to incentivize socially-optimal effort. But since mandating suboptimal effort via such an overreaction is undesirable, we consider a simpler problem setting: suppose that the regulator has access to the platform's proprietary model and its underlying data, which can now be used as the public model robust to performativity. The regulator thus has knowledge of the harm function h. Suppose further that the societal costs of harm D are also given; the only missing information is the cost function c, or the platform's costs of effort to mitigate disinformation.

Proposition 5 *There is no way of adjusting the effort e_c required by the public model, purely as a function of the harm function h and the cost of damages D, without*

regard to the cost function c, such that the platform is always incentivized to exert the socially-optimal level of effort.

Proof. Consider the simplest possible case where we assume there exist only two possible cost functions, c_1 and c_2. Let e_1^* be the socially-optimal effort induced by cost function c_1 and e_2^* be that induced by cost function c_2, and let $e_2^* > e_1^*$. Note that the e_1^* and e_2^* can be trivially computed from equation (3.1) by equating the marginal social welfare to zero.

Suppose that the actual socially-optimal effort is e_1^*. Note first that if the regulator sets the public model to require effort $e_c = e_1^*$, there should not be any increase in this level of effort because a transition to some new $e_h > e_c$ will mandate excessive and therefore suboptimal effort as the platform at least follows the public model's specified effort (by assumption). Thus, if the public model is set to require effort e_1^*, we are done.

However, now suppose that the actual socially-optimal effort is e_2^* and the public model currently specifies e_1^* as the required effort. Then, an increase in effort to e_2^* is necessary to incentivize the platform to exert the socially-optimal effort (contradiction).

Similarly, a decrease in the effort required by the public model does not guarantee that platform is always induced to exert the socially-optimal effort level: If e_2^* is socially-optimal, then a decrease in the required effort to some $e_c < e_2^*$ does not guarantee that the platform will continue to exert effort at e_2^*; the platform may exert less and therefore socially suboptimal effort at e_c, since without increasing the required effort, there is no way for a static public model to induce the platform to exert more than the prescribed effort as shown in Proposition 1. Therefore, if the public model is set to require e_1^*, the platform is no longer guaranteed to exert the socially-optimal e_2^*.

Thus, because no adjustment to the public model's required effort works, the regu-

lator needs to know whether the platform's true cost function is c_1 or c_2 to incentivize the socially-optimal effort at all times. And since there exist more than just two possible choices for the platform's actual cost function, there is no way for a regulator to guarantee that the platform exerts socially optimal effort for mitigating disinformation with any increasing or decreasing adjustments to the public model's required effort. ■

This result shows that even if a regulator has precise control over the public model, without knowledge of the platform's costs, there is no way to set up the public model's effort threshold such that the platform's individually-optimal effort level is always socially optimal. Thus, since social platforms' costs of precautionary effort underpin the incentive problem, it is crucial to model these costs in more detail to better understand their incentives relating to the control of disinformation. Determining how engagement translates to money, therefore, serves as an important avenue for future exploration, as platforms risk losing out on engagement revenue with content moderation.

Chapter 5

Takeaways and Prescriptions

The focus of this chapter is to emphasize areas where this thesis work may be extended for more fruitful insights regarding the regulation of social platforms. We briefly restate our main assumptions and results before segueing into the main discussion.

Our formal results captured in the previous chapter follow from certain key assumptions supporting our descriptive model. To restate concisely, we assume that only social platforms have the best access to the technical resources, expertise, and user data required to effectively mitigate disinformation. A regulator can only imperfectly monitor a social platform's effort to reduce the spread of toxic content, based on some public notion of harmfulness of content that specifies if certain content ought not to be hosted by the platform. However, disinformation authors can inundate platforms with new forms of harmful content. As such, the public standard for content moderation will prove to be ineffectual at identifying rapidly evolving forms of disinformation, unless some harmful event occurs, thereby enabling the public to update their beliefs about content that is harmful and ought to be removed from platforms.

Under these conditions, Theorem 1 and Proposition 4 demonstrate that the public standard of effort required to mitigate disinformation — in terms of specifying what content must be limited from platforms — must become excessive, or socially suboptimal, as a response to the occurrence of some harm, in order to induce a social platform to perform the adequate, or socially optimal, level of control of disinforma-

tion. Clearly these results exhibit undesirable properties; regulation of platforms via mandating excessive content moderation is not a practical recommendation. Furthermore, our impossibility result (Proposition 5) captures another undesirable property: even if a regulator possesses the same technical expertise and resources as a social platform, there is no way to induce the platform to control disinformation adequately via our mode of ex ante negligence regulation, without knowledge of the platform's costs of content moderation efforts.

Despite these perhaps unenviable conclusions, our modeling exercise offers valuable insights into the incentive issues relevant to platforms' control of online disinformation. Moreover, our results provide a lens through which further regulatory prescriptions for controlling disinformation might be derived. Accordingly, we highlight our modeling constraints and assumptions that will be worthwhile to relax or expand upon in derivative work.

5.1 Homogeneous Harm

One of the primary features of our model is expressing the harm from disinformation as a binary event and assuming that it effects all people equally. Recall our assumption from Chapter 3 that the platform possesses a proprietary machine learning model that assigns for every item of content the probability that it will cause some harmful event H yielding a fixed societal cost D. Essentially, such a model considers only one particular dimension of harm for every content item, that is, how likely is it for some event to occur due to that content being hosted and its induced engagement. This binary notion of harm might seem restrictive, especially because the harm from disinformation can manifest in many forms: rare events such as the Capitol Hill riots or the Pizzagate shooting are dramatic and immediately observable, in comparison to harm from the degradation of public discourse or from the spread of climate change denial or anti-vaccine propaganda, which are more subtle manifestations. Regardless, our binary notion of harm is without loss of generality.

To illustrate, recall that we consider a simple probability density function $h(e) = \Pr[H|e]$ of a harmful event for a certain level of mitigating effort e, and think of this event as yielding a fixed cost to society, D. Indeed, equation (3.1) captures the expected amount of harm $h(e)D$ in terms of quantifying the expected societal costs should the harmful event occur. But different levels of mitigating effort might give rise to different types of harm. Furthermore, there might exist multiple dimensions of harmfulness associated with the different kinds of user-generated content. For example, consider a model that simply measures the amount of toxicity in a content item through explicitly checking its text, image or video message [45], and not, as in our setup, the likelihood of that item causing a specific harmful event. Nonetheless, equation (3.1) can be augmented to capture different types of harm: we will simply substitute our harm function with different probability distribution functions for the different dimensions of harm and include the associated societal costs. This practice will preserve our model's notion of quantifying expected harm. Thus, it is straightforward to extend our model for different kinds of harm.

5.2 Heterogeneous Content

While our notion of measuring content harmfulness via a binary harmful event is without loss of generality, it is meaningfully different to consider the heterogeneity of content in terms of how harmful a particular piece of content is and how much benefit it brings to a social platform. Our simple model of a platform's costs as a function of effort, $c(e)$, implicitly encodes the platform's valuation for content: recall that $c(e)$ comprises the indirect costs of the platform losing out on advertising revenue due to the loss of user engagement with the deployment of content moderation efforts; moreover, our model implies homogeneity of all content with respect to the amount of user engagement for each item. In reality, however, just as content is not homogeneous in terms of the varying degrees of harmfulness of each item, content will also differ in the levels of user engagement attained. Therefore, modeling this heterogeneous

relationship of the harm and benefit of content will likely drive different conclusions.

For example, with such explicit modeling, one question we might hope to answer is whether highly toxic content is more likely to produce high levels of user engagement (in the form of likes, shares, retweets, comments etc.) than less toxic content, thereby being more valuable to the platform. Recent examples indicate the prevalence of this phenomenon [3, 4, 6, 19, 46]. In theory, assuming it could only pick one, a social platform would prefer hosting a content item that attains or is predicted to attain more user engagement instead of one that does not attain as much, all else being equal [2, 12]. But this approach of maximizing user engagement becomes problematic when content that is highly toxic, false, or misleading achieves high levels of user engagement or virality, since the costs of harmful content are borne exclusively by society. Thus, answering our question will shed light on the degree to which the incentives of social platforms relating to the control of disinformation are misaligned with those of society.

Additionally, understanding the relationship between toxicity and benefit of content is important as it informs the nature of any regulatory restrictions required to realign these incentives. We may find instances where harmful content only attains low levels of user engagement on social platforms. The regulation mechanisms to induce prompt mitigation of disinformation will likely differ in this situation in comparison to the one described previously. Because the regulator is social welfare-maximizing, we care about the overall costs of the regulatory mechanisms imposed on social platforms. As such, a platform should not be needlessly penalized for hosting viral content, especially if it is the case that highly toxic content is less likely to attain high levels of user engagement.

In essence, a social platform benefits from more user engagement than less, irrespective of whether such engagement is induced from harmful or benign content. But because it is only society that incurs the costs of harmful content, the idea behind any regulatory scheme is to reduce the amount of toxicity spread on platforms. A

regulator cannot directly exert the effort required to contain toxic content without access to the platform's proprietary data and technical resources. Thus, a possible approach to regulate platforms might be via devising a mechanism based on taxation.

5.2.1 Taxing Toxicity

A *Pigouvian tax* is a tax on a market transaction that generates a negative externality borne by individuals not directly involved in the transaction [47]. Such a tax may be levied on firms that generate pollution as a result of their production processes. The carbon tax is one such example of a Pigouvian tax: when individuals purchase goods produced by firms, carbon emissions might be a by-product of the production process causing air pollution, which is a negative externality impacting the environment and imposing a cost on individuals not directly involved in the initial production and purchase transaction. Thus, a regulatory body specifies a price that firms must pay for the amount of carbon they emit. This tax is meant to "internalize" the costs of the externality to the firm's production process.

Social platforms offer products and services enabling a multitude of users to connect with each other, and create and share content. But while most of these services are free for users, the actual customers of platforms are advertisers willing to purchase advertising space on the platforms. Essentially, the more users a platform has, the more lucrative it is for advertisers to pay for the platform's services to target them with ads; moreover, the more time these users spend engaging with other users and content on the platform, the greater the opportunity for the platform to cater to the precise needs of advertisers. Note that in this particular transaction, a negative externality may be generated from user engagement that is induced from toxic content on the platform: such engagement imposes a cost only on the users of social media, or society overall, and not on the platform and advertisers involved in the original transaction.

Thus, social platforms exhibit the precise criterion of generating negative exter-

nalities that calls for the levying of a Pigouvian tax. To elaborate, consider a toy example. Suppose a platform's true valuation for some content is $1M calculated in terms of the engagement and eyeballs it attracts, thereby affording the platform more opportunity to sell advertising space. Now further suppose that a subset of this content is toxic, and the cost of damages to society due to the externality generated from user engagement are estimated to be at most $150K. From the lens of welfare economics, a tax should not be imposed on the platform for hosting such content due to the positive net social benefit. However, if instead the two costs are swapped — the cost of the externality exceeds the platform's valuation for content — then a tax should be levied on the platform if it chooses to host this content. Under ideal circumstances, the amount of tax would be equal to the cost of damages due to the externality generated [47].

Naturally, a regulator will require a good predictive harm model to measure content harmfulness, or the extent to which a piece of content will give rise to toxic engagement. An efficient taxation scheme will pick a tax rate proportional to the measure of toxicity of content; that is, content classified as ludicrously toxic should have the highest tax rate, while content classified as acceptable the lowest, assuming more toxic content hosted online makes the occurrence of a harmful event more likely. Ideally, the hope is that with a good harm model, the regulator can restrict taxation to the subset of content that drives toxic engagement, while leaving other benign content untaxed; and furthermore, an increasing tax rate with the toxicity of content will induce the platforms to take proactive measures to limit the spread of such content in order to avoid paying large sums of taxes.

However, devising a taxation scheme as such might not be practical. Developing a good harm model to measure content toxicity will likely require an access to data and expertise that is only available to platforms. Therefore, a mechanism designer (regulator) might instead impose taxation in a more crude manner, for example, by levying a tax on user engagement on the platform more generally, rather than

on the harmfulness of hosted content. As discussed previously, although regulating engagement does not bear directly on content toxicity — the entity we wish to control on social platforms — it might nevertheless be the only means of controlling online disinformation through taxation. In terms of implementation, the regulator can ask the platform report its cost function for moderating content and then tax the platform based on its report. Thus, this crude notion of taxation fits well with our simple model of the cost function, as the indirect costs of effort essentially capture the value of engagement. Furthermore, as shown in Proposition 5, the platform's costs of effort underpin the incentive problem for disinformation mitigation; thus, any effective mechanism must in some way be responsive to these costs. Naturally, such a mechanism must also factor in incentives that might prevent the platform from misreporting its true cost function for moderating content, in the hopes of attaining a lower tax rate, for instance.

5.3 Prospective Versus Retrospective Harm

Throughout this work, we have only considered a *prospective*, or predictive, notion of harm. That is, our simple harm model predicts the likelihood of a unique harmful event for a given level of mitigating effort. In Chapter 3, however, through our discussion of the strict liability standard, we are effectively dealing with a *retrospective* notion of harm: we claim that regulatory enforcement of limiting disinformation via a strict liability standard necessitates ascribing responsibility for harm after the fact. The challenge of assigning blame lies in identifying the direct causal links between the harmful event and a specific social platform, given the presence of a multitude of platforms, the inter-connectivity of users and thus also content shared on the platforms, and harm being carried out by individuals. Thus, in our assertion about the impracticality of a strict liability standard, we implicitly assume the difficulty of programming a retrospective model of harm, that is, one that can correctly ascribe responsibility to a platform for harm.

Indeed, it is infeasible to expect a regulator to have a perfect retrospective harm model programmed to assign blame for harm. Because if such a model did exist, without making any claims about its exact form, it is clear from our analysis that a regulator could simply enforce a strict liability standard to penalize the platform if it were deemed guilty — via use of this retrospective harm model — for causing some harmful event, thereby effectively aligning the incentives of social platforms with that of society. However, one question to explore is whether we can utilize a softer notion of retrospective reasoning in combination with our prospective harm model to regulate disinformation. Concretely, is there some combination of a predictive harm model and a model short of fully attributing responsibility to social platforms for harm after the fact such that a regulator can incentivize the optimal effort for mitigating disinformation?

Exploring the incentive properties of regulation via a prospective harm model complemented by retrospective reasoning can be a worthwhile future extension of our work. We have already shown that a perfect prospective model of harm is insufficient to incentivize adequate control of disinformation: indeed, our impossibility result (Proposition 5) demonstrates that even if a regulator possesses a perfect predictive model of harm, knowledge of the platform's costs of effort is necessary for any hope of inducing the socially-optimal level of effort from the platform. Incorporating a retrospective notion of harm can, hence, complement a predictive harm model for detecting disinformation. There is already a sense in which a good predictive model of harm relies on retrospective reasoning: Donald Trump's misinformed tweet about the efficacy of hydroxychloroquine in curing Covid-19 drove purchases of this anti-malarial drug and increased false rumours surrounding the pandemic [48, 49]; but this fact offers an opportunity to program classification models with more information to promptly flag future iterations of such misinformation.

Ultimately, retrospective reasoning is beneficial as such because it can elucidate new data on which a good predictive model of harm should be trained. We can ask

questions about what ground truth data should be included to train a good predictive model of harm, or about why it would make more sense to disregard previous training data in light of new information post hoc, for example. Furthermore, unlike a predictive harm model, a retrospective model cannot be gamed and circumvented by bad actors, for it relies on determining causes of harm after the fact. The practice of determining causes for events, in fact, is related to the concept of *actual causality*.

Actual Causality Halpern [50] describes actual causality as the problem of determining what specific events explain a particular observation or incident about the world. This notion is contrasted to that of general causality, or *type causality*, which is forward-looking and used for predictions. To elaborate, actual causality is backward-looking in the sense that we know a particular event to have occurred and seek out explanations for why it did; type causality, conversely, focuses on general causal statements like "excessive alcohol assumption causes liver failure" [50]. Thus, in the context of societal harm from disinformation, the problem of actual causality looks at a specific harmful incident, say the Capitol Hill riots, and asks what events in particular caused it to occur. These factors could include individuals from certain conspiracy groups, which in turn might have been nudged by particular pieces of content spread on social platforms, which in turn were the means by which individuals coordinated the attacks [9].

Thus, programming a retrospective harm model for disinformation implies solving the problem of actual causality: in order to assign responsibility for some harmful event after the fact. But while solving actual causality perfectly might be intractable, there is value to be derived from considering a weaker notion of causality in the form of retrospective reasoning, as discussed previously. Consequently, a promising future direction of this will explore the merits of assigning a weaker form of liability to platforms, in conjunction with negligence, both ex ante and ex post, in order to target a better regulatory response for controlling disinformation on social platforms

— especially in comparison to an overreaction, or of such response being impossible under our descriptive model's constraints.

Chapter 6

Conclusion

Events like the Covid-19 "infodemic" or the Capitol Hill riots are recent examples of the harm associated with disinformation. There is increasing evidence that the failure of social media platforms to control the spread of disinformation is due to incentive issues rather than a lack of technical ability [12, 14, 46]. This work provides a formal analysis of these incentive issues that adapts the standard principal-agent framework to incorporate the unique features of the domain. Our formal model, although stylized, includes what we take to be key aspects of the setting, including the performativity of disinformation classification and public backlash as a response to harmful events. Our formal results provide insights for the effective regulation of social media platforms.

We argue that although a strict liability standard would theoretically align the platform's incentives with those of society, it is unlikely to be practical given the difficulty of assigning responsibility for harmful events to specific instances of disinformation ex post. Using our formal model, we derive a number of results relating to the use of a negligence standard. Most importantly, we show that in the absence of a public backlash to harmful events, there is no monitoring scheme that can induce platforms to perform a socially-optimal level of control of disinformation. However, when the public model includes the possibility of overreacting to a harmful event by requiring a greater than socially optimal level of effort, a platform can be incentivized to exert

more diligent effort than explicitly required by the regulator. Ultimately, because mandating suboptimal effort is undesirable, we advocate devising mechanisms that may elicit platforms' costs of precautionary effort for limiting disinformation. Our final impossibility result emphasizes this call for transparency as we show that absent knowledge of these costs, there is no way to reliably induce a socially-optimal level of control of disinformation under a negligence standard, even if all other parameters of the setting are given.

Our model makes a number of simplifying assumptions. Treating public standards as an explicit model implies that a platform can guarantee a given probability of escaping punishment if it conforms to an explicit standard, which is an oversimplification of reality. The assumption that the platform can perfectly tune its model is also unrealistic; technical challenges, although they may not pose the main obstacle to the practical control of disinformation, are nevertheless a real issue [2]. Extending the model to more richly model these aspects are important directions for future work, in addition to those explored in Chapter 5.

Disinformation is one of the most urgent problems facing society. But it is a problem driven by incentives as much as by technology. This work takes a first step toward explicitly modeling the incentive issues that must be accounted for by any effective solution to the problem.

6.1 Ethical Considerations

Regulating social media is an especially sensitive issue. Although allowing disinformation to spread unchecked is clearly unsustainable, disinformation control always runs the risk of becoming censorship. In this work, we take the existence of a "public model" of acceptable postings for granted. However, the content of this public standard is a question of societal standards that can be settled only by public debate. Similarly, we analyze the use of "monitoring" without specifying its exact form. A naively implemented monitoring scheme would run the risk of serious privacy viola-

tions.

References

[1] S. Kumar and N. Shah, *False information on web and social media: A survey*, 2018. arXiv: 1804.08559 [cs.SI].

[2] K. Hao, *He got facebook hooked on ai. now he can't fix its misinformation addiction*, Available at https://www.technologyreview.com/2021/03/11/1020600/facebook-responsible-ai-misinformation/, Mar. 2021.

[3] *Anti-racism protests: Divisive disinformation narratives go viral on facebook, racking up over 26 million estimated views*, Available at https://secure.avaaz.org/campaign/en/anti_protest_disinformation/, Jun. 2020.

[4] C. Silverman, *This analysis shows how viral fake election news stories outperformed real news on facebook*, Available at https://www.buzzfeednews.com/article/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook, Nov. 2016.

[5] *Managing the covid-19 infodemic: Promoting healthy behaviours and mitigating the harm from misinformation and disinformation*, Available at https://www.who.int/news/item/23-09-2020-managing-the-covid-19-infodemic-promoting-healthy-behaviours-and-mitigating-the-harm-from-misinformation-and-disinformation, Sep. 2020.

[6] *How facebook can flatten the curve of the coronavirus infodemic*, Available at https://secure.avaaz.org/campaign/en/facebook_coronavirus_misinformation/, Apr. 2020.

[7] T. Telford, *Anti-vaxxers are spreading conspiracy theories on facebook, and the company is struggling to stop them*, Available at https://www.washingtonpost.com/business/2019/02/13/anti-vaxxers-are-spreading-conspiracy-theories-facebook-company-is-struggling-stop-them/, Feb. 2019.

[8] *Facebook: From election to insurrection*, Available at https://secure.avaaz.org/campaign/en/facebook_election_insurrection/, Mar. 2021.

[9] *Capitol attack was months in the making on facebook*, Available at https://www.techtransparencyproject.org/articles/capitol-attack-was-months-making-facebook, Jan. 2021.

[10] J. W. C. Marc Fisher and P. Hermann, *Pizzagate: From rumor, to hashtag, to gunfire in d.c.* Available at https://www.washingtonpost.com/local/pizzagate-from-rumor-to-hashtag-to-gunfire-in-dc/2016/12/06/4c7def50-bbd4-11e6-94ac-3d324840106c_story.html, Dec. 2016.

[11] A. Stevenson, *Facebook admits it was used to incite violence in myanmar*, Available at https://www.nytimes.com/2018/11/06/technology/myanmar-facebook.html, Nov. 2018.

[12] M. I. Kevin Roose and S. Frenkel, *Facebook struggles to balance civility and growth*, Available at https://www.nytimes.com/2020/11/24/technology/facebook-election-misinformation.html, Nov. 2020.

[13] C. Silverman and R. Mac, *"facebook gets paid"*, Available at https://www.buzzfeednews.com/article/craigsilverman/facebook-ad-scams-revenue-china-tiktok-vietnam, Dec. 2020.

[14] D. K. Citron and M. A. Franks, "The internet as a speech machine and other myths confounding section 230 reform," *Boston University School of Law, Public Law and Legal Theory Paper Series*, 2020.

[15] X. Zhou and R. Zafarani, "A survey of fake news: Fundamental theories, detection methods, and opportunities," *ACM Comput. Surv.*, vol. 53, no. 5, Sep. 2020, ISSN: 0360-0300. DOI: 10.1145/3395046. [Online]. Available: https://doi.org/10.1145/3395046.

[16] A. Mosseri, *Addressing hoaxes and fake news*, Available at https://about.fb.com/news/2016/12/news-feed-fyi-addressing-hoaxes-and-fake-news/, Dec. 2016.

[17] K. Coleman, *Introducing birdwatch, a community-based approach to misinformation*, Available at https://blog.twitter.com/en_us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation, Jan. 2021.

[18] P. Nakov *et al.*, *Automated fact-checking for assisting human fact-checkers*, 2021. arXiv: 2103.07769 [cs.AI].

[19] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, pp. 1146–1151, Mar. 2018. DOI: 10.1126/science.aap9559.

[20] H. F. Ryan Dansby and H. Ma, *Ai advances to better detect hate speech*, Available at https://ai.facebook.com/blog/ai-advances-to-better-detect-hate-speech, May 2020.

[21] K. Kumar and G. Gopalan, "Detecting misinformation in online social networks using cognitive psychology," *Human-centric Computing and Information Sciences*, vol. 4, p. 14, Sep. 2014. DOI: 10.1186/s13673-014-0014-x.

[22] *"facts before rumors" campaign just began by the ibs data science group*, Available at https://www.ibs.re.kr/cop/bbs/BBSMSTR_000000000739/selectBoardArticle.do?nttId=18241, Mar. 2020.

[23] *White paper: Correcting the record*, Available at https://secure.avaaz.org/campaign/en/correct_the_record_study/, Apr. 2020.

[24] P. Dütting, T. Roughgarden, and I. Talgam-Cohen, "Simple versus optimal contracts," *CoRR*, vol. abs/1808.03713, 2018. arXiv: 1808.03713. [Online]. Available: http://arxiv.org/abs/1808.03713.

[25] P. Duetting, T. Roughgarden, and I. Talgam-Cohen, "The complexity of contracts," *CoRR*, vol. abs/2002.12034, 2020. arXiv: 2002.12034. [Online]. Available: https://arxiv.org/abs/2002.12034.

[26] P. Duetting, T. Ezra, M. Feldman, and T. Kesselheim, *Combinatorial contracts*, 2021. arXiv: 2109.14260 [cs.GT].

[27] S. J. Grossman and O. D. Hart, "An analysis of the principal-agent problem," *Econometrica*, vol. 51, no. 1, pp. 7–45, 1983, ISSN: 00129682, 14680262. [Online]. Available: http://www.jstor.org/stable/1912246.

[28] R. Innes, "Optimal liability with stochastic harms, judgement-proof injurers, and asymmetric information," *International Review of Law and Economics*, vol. 19, no. 2, pp. 181–203, 1999, ISSN: 0144-8188. DOI: https://doi.org/10.1016/S0144-8188(99)00004-6. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0144818899000046.

[29] M. A. Cohen, "Optimal enforcement strategy to prevent oil spills: An application of a principal-agent model with moral hazard," *Journal of Law & Economics*, vol. 30, p. 23, 1987.

[30] S. Shavell, "Risk sharing and incentives in the principal and agent relationship," *The Bell Journal of Economics*, vol. 10, no. 1, pp. 55–73, 1979, ISSN: 0361915X, 23263032. [Online]. Available: http://www.jstor.org/stable/3003319.

[31] A. Alaphilippe, *Disinformation is evolving to move under the radar*, Available at https://www.brookings.edu/techstream/disinformation-is-evolving-to-move-under-the-radar/, Feb. 2021.

[32] S. Baliga and E. Maskin, "Chapter 7 - mechanism design for the environment," in *Environmental Degradation and Institutional Responses*, ser. Handbook of Environmental Economics, K.-G. Mäler and J. R. Vincent, Eds., vol. 1, Elsevier, 2003, pp. 305–324. DOI: https://doi.org/10.1016/S1574-0099(03)01012-X. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S157400990301012X.

[33] R. Innes, "Enforcement costs, optimal sanctions, and the choice between ex-post liability and ex-ante regulation," *International Review of Law and Economics*, vol. 24, no. 1, pp. 29–48, 2004, ISSN: 0144-8188. DOI: https://doi.org/10.1016/j.irle.2004.03.003. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0144818804000183.

[34] J. C. Teitelbaum, "A unilateral accident model under ambiguity," *The Journal of Legal Studies*, vol. 36, no. 2, pp. 431–477, 2007, ISSN: 00472530, 15375366. [Online]. Available: http://www.jstor.org/stable/10.1086/511895.

[35] K. Franek, *How facebook makes money: Business model explained*, Available at https://www.kamilfranek.com/how-facebook-makes-money-business-model-explained/, Apr. 2021.

[36] S. Rodriguez, *Sandberg says u.s. capitol riot was 'largely' not organized on facebook*, Available at https://www.cnbc.com/2021/01/11/sandberg-says-us-capitol-riot-was-not-organized-on-facebook.html, Jan. 2021.

[37] *Community standards enforcement report*, Available at https://transparency.facebook.com/community-standards-enforcement, Feb. 2021.

[38] J. Perdomo, T. Zrnic, C. Mendler-Dünner, and M. Hardt, "Performative prediction," in *Proceedings of the 37th International Conference on Machine Learning*, H. D. III and A. Singh, Eds., ser. Proceedings of Machine Learning Research, vol. 119, PMLR, Jul. 2020, pp. 7599–7609. [Online]. Available: http://proceedings.mlr.press/v119/perdomo20a.html.

[39] X. Zhou, A. Jain, V. V. Phoha, and R. Zafarani, "Fake news early detection: A theory-driven model," *Digital Threats: Research and Practice*, vol. 1, no. 2, pp. 1–25, 2020.

[40] J. Y. Khan, M. T. I. Khondaker, S. Afroz, G. Uddin, and A. Iqbal, "A benchmark study of machine learning models for online fake news detection," *Machine Learning with Applications*, vol. 4, p. 100 032, Jun. 2021, ISSN: 2666-8270. DOI: 10.1016/j.mlwa.2021.100032. [Online]. Available: http://dx.doi.org/10.1016/j.mlwa.2021.100032.

[41] M. Bastos and S. T. Walker, *Facebook's data lockdown is a disaster for academic researchers*, Available at https://theconversation.com/facebooks-data-lockdown-is-a-disaster-for-academic-researchers-94533, Apr. 2018.

[42] B. Booker, *Facebook removes 'stop the steal' content; twitter suspends qanon accounts*, Available at https://www.npr.org/sections/insurrection-at-the-capitol/2021/01/12/956003580/facebook-removes-stop-the-steal-content-twitter-suspends-qanon-accounts, Jan. 2021.

[43] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, Second. The MIT Press, 2018.

[44] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004. DOI: 10.1017/CBO9780511804441.

[45] A. Sheth, V. L. Shalin, and U. Kursuncu, *Defining and detecting toxicity on social media: Context and knowledge are key*, 2021. arXiv: 2104.10788 [cs.SI].

[46] A. Cranz and R. Brandom, *Facebook encourages hate speech for profit, says whistleblower*, Available at https://www.theverge.com/2021/10/3/22707860/facebook-whistleblower-leaked-documents-files-regulation, Oct. 2021.

[47] *Pigouvian tax definition*, https://taxfoundation.org/tax-basics/pigouvian-tax/, Accessed: 2022-01-20.

[48] A. C. Estes, *Hydroxychloroquine conspiracies are back, but trump's the patient now*, Available at https://www.vox.com/recode/2020/10/7/21504748/hydroxychloroquine-trump-covid-treatment-misinformation, Oct. 2020.

[49] M. R. Haupt, J. Li, and T. K. Mackey, "Identifying and characterizing scientific authority-related misinformation discourse about hydroxychloroquine on twitter using unsupervised machine learning," *Big Data & Society*, vol. 8, no. 1, p. 20 539 517 211 013 843, 2021. DOI: 10.1177/20539517211013843. eprint: https://doi.org/10.1177/20539517211013843. [Online]. Available: https://doi.org/10.1177/20539517211013843.

[50] J. Y. Halpern, *Actual Causality*. The MIT Press, 2016, ISBN: 0262035022.