This article was downloaded by: [193.60.238.99] On: 07 April 2024, At: 08:14
Publisher: Institute for Operations Research and the Management Sciences (INFORMS)
INFORMS is located in Maryland, USA

Strategy Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Soft Governance Across Digital Platforms Using Transparency

Anil R. Doshi, William Schmidt

To cite this article:
Anil R. Doshi, William Schmidt (2024) Soft Governance Across Digital Platforms Using Transparency. Strategy Science

Published online in Articles in Advance 13 Mar 2024

. https://doi.org/10.1287/stsc.2023.0006

Full terms and conditions of use: https://pubsonline.informs.org/Publications/Librarians-Portal/PubsOnLine-Terms-and-Conditions

This article may be used only for the purposes of research, teaching, and/or private study. Commercial use or systematic downloading (by robots or other automatic processes) is prohibited without explicit Publisher approval, unless otherwise noted. For more information, contact permissions@informs.org.

The Publisher does not warrant or guarantee the article's accuracy, completeness, merchantability, fitness for a particular purpose, or non-infringement. Descriptions of, or references to, products or publications, or inclusion of an advertisement in this article, neither constitutes nor implies a guarantee, endorsement, or support of claims made of that product, publication, or service.

Copyright © 2024 The Author(s)

Please scroll down for article—it is on subsequent pages

With 12,500 members from nearly 90 countries, INFORMS is the largest international association of operations research (O.R.) and analytics professionals and students. INFORMS provides unique networking and learning opportunities for individual professionals, and organizations of all types and sizes, to better understand and use O.R. and analytics tools and methods to transform strategic visions and achieve better outcomes.
For more information on INFORMS, its publications, membership, or meetings visit http://www.informs.org

https://pubsonline.informs.org/journal/stsc

STRATEGY SCIENCE
Articles in Advance, pp. 1–20
ISSN 2333-2050 (print), ISSN 2333-2077 (online)

Downloaded from informs.org by [193.60.238.99] on 07 April 2024, at 08:14 . For personal use only, all rights reserved.

Soft Governance Across Digital Platforms Using Transparency

Anil R. Doshi,[a,*] **William Schmidt**[b]

[a] UCL School of Management, London E14 5AA, United Kingdom; [b] Goizueta Business School, Emory University, Atlanta, Georgia 30322
*Corresponding author

Contact: anil.doshi@ucl.ac.uk, https://orcid.org/0000-0002-8489-3373 (ARD); william.schmidt@emory.edu, https://orcid.org/0000-0003-1424-7437 (WS)

Received: January 5, 2023
Revised: November 4, 2023; January 3, 2024
Accepted: February 13, 2024
Published Online in Articles in Advance: March 13, 2024

https://doi.org/10.1287/stsc.2023.0006

Copyright: © 2024 The Author(s)

Abstract. Platform governance helps align the activities of participating actors to deliver value within the platforms. These platforms can operate in environments where governance is intentionally or conventionally weak in favor of open access, frictionless transactions, or free speech. Such low- or no-governance environments leave room for illegitimate actors to penetrate platforms with illegitimate content or transactions. We propose that an external observer can employ transparency mechanisms to establish "soft" governance that allows participants in a low-governance environment to distinguish between sources of legitimate and illegitimate content. We examine how this might work in the context of disinformation Internet domains by training a machine learning classifier to discern between low-legitimacy from high-legitimacy content providers based on website registration data. The results suggest that an independent observer can employ such a classifier to provide an early, although imperfect, signal of whether a website is intended to host illegitimate content. We show that the independent observer can be effective at serving multiple platforms by providing intermediate prediction results that platforms can align with their unique governance priorities. We expand our analysis with a signaling game model to ascertain whether such a soft governance structure can be resilient to adversarial responses.

Open Access Statement: This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License. You are free to download this work and share with others, but cannot change in any way or use commercially without permission, and you must attribute this work as "*Strategy Science*. Copyright © 2024 The Author(s). https://doi.org/10.1287/stsc.2023.0006, used under a Creative Commons Attribution License: https://creativecommons.org/licenses/by-nc-nd/4.0/."

Funding: Funding for this research was provided by UCL School of Management and Emory University.

Supplemental Material: The online appendix is available at https://doi.org/10.1287/stsc.2023.0006.

Keywords: platform governance • soft governance • transparency • disinformation

1. Introduction

Platform research and practical applications have grown as organizations and individuals become increasingly connected across digital environments (Adner et al. 2019). Platforms "function as an interface between different groups of users and facilitate value-creation exchanges" (Cennamo and Santalo 2013, p. 1331). These different groups can have diverse and sometimes diverging interests. This has motivated a growing body of research on platform governance and the governance policies that platforms may adopt (Tiwana et al. 2010, Koo and Eesley 2020, Rietveld et al. 2021).

However, even well-governed platforms may be vulnerable if they operate in low-governance environments. Such environments may lack mechanisms and policies to regulate the quality of the actors in the environment. While low-governance environments can serve as a conduit for exchanges of legitimate content and transactions among platforms and external actors, they leave open the possibility that illegitimate actors can establish illegitimate content and transactions in these environments and introduce them into otherwise well-governed platforms.

To mitigate the threat to platforms posed by illegitimate actors operating in low-governance environments, we examine the platform in the context of its broader governance environment and propose how an independent observer might improve that environment, even if it lacks formal enforcement authority. We start by describing three levels of governance that affect platforms and environments: platforms, governments, and regulatory intermediaries. Platforms establish the structure and rules for platform participation and provide the primary form of governance within the boundaries of a platform; governments handle criminal and civil regulation within the jurisdiction(s) that the platform operates and can add a layer of support to platform-spanning transactions; and regulatory intermediaries provide targeted governance functions in the broader environment. Using this

Downloaded from informs.org by [193.60.238.99] on 07 April 2024, at 08:14 . For personal use only, all rights reserved.

framework, we evaluate how an independent observer acting as a regulatory intermediary can mitigate the impact of illegitimate actors by collecting and assessing identity transparency (hereafter, transparency) signals from content providers. This assessment allows the observer to exercise a form of soft governance, which we define as indirect influence without formal authority that elicits positive behavior from participants, especially in a low- or no-governance environment. While we treat the observer as an independent entity, it could represent a function within an existing entity, including Internet security companies, independent researchers, or government agencies. We use transparency as a measure of how easy it is to identify the actors responsible for content or transactions in an environment. A similar construct has been employed in other settings to mitigate different forms of illegitimate and illicit activity, including discouraging online sales of stolen goods (Schakowsky 2021) and human trafficking in supply chains (Harris 2015).

Intuitively, being transparent is undesirable for bad actors because it facilitates proper attribution of their behavior, which allows authorities to impose costs on those bad actors. Kevin Mandia, Chief Executive Officer of the Internet security firm FireEye, Inc., makes this point clear in the context of disinformation during his testimony to the Select Committee on Intelligence of the U.S. Senate under the title "Disinformation: A Primer in Russian Active Measures and Influence Campaigns." He states, "When you have attribution right, then you can consider the proportional response and the other tools at your disposal as diplomats to make sure we have the deterrence we need" (S.H. 115-40 2017). Another example of how transparency is costly to bad actors is the Integrity, Notification, and Fairness in Online Retail Marketplaces for Consumers (INFORM Consumers) Act, which requires online resellers to authenticate their identity as a deterrent to illicit trade. As reported in *The New York Times*, "By forcing organized retail criminals to verify who they are, the bill would discourage them from selling and thus exposing themselves to prosecution" Coy (2021).

To illustrate this mechanism, we use the context of online disinformation and consider how an external observer using transparency signals can mitigate the spread of disinformation. We propose that the external observer can leverage existing domain registration data to detect disinformation domains near the time of a domain's registration. The external observer's capacity for soft governance stems from its ability to credibly, albeit incompletely, identify disinformation domains and share the details of those domains with others for further validation, including researchers, Internet security firms, social media platforms, and media outlets. This additional validation is useful because the initial identification may be imperfect.

Although domain registration data are publicly accessible, it has little to no explicit information about the intended use of the domain. We provide a practical validation that a reasonably accurate signal of disinformation domains can be detected by training a machine learning classifier using the initial registration data from a data set of known disinformation domains and a random sample of general domains. Even with limited data and a standard machine learning algorithm, the predictive results using a holdout sample of data are encouraging. We present the model's predictive performance using the area under the curve (AUC) for the classifier's precision-recall (PR) curve. This measure is bound between 0 (no predictive power) and 1 (perfect predictive power). A PR curve is a common measure of prediction performance in settings such as ours with strong class imbalance. Our base model yields a PR AUC of 0.521. This result exceeds the performance of a random classifier, which yields a PR AUC of 0.112 (equivalent to the proportion of disinformation domains in the data used to assess the performance of the model). Over a series of robustness checks, we produce alternative classifiers with PR AUCs ranging from 0.502 to 0.615. To better establish that the prediction results are based on identity transparency signals, we train a classifier using only a subset of registration features related to the registrant's identity. The model's predictive performance is nearly equivalent to that obtained by training on features drawn from all the registration data. The results demonstrate that seemingly innocuous disclosures can contain transparency signals that are detectable by a machine.

The observer's analysis can be leveraged by platforms to support their platform-centric efforts to suppress disinformation. Platforms can have different governance policies regarding disinformation (Kuan and Lee 2023), which can manifest in a different tolerance for false positive and false negative predictions. We confirm that the observer's analysis can better align with the needs of platforms' disparate governance policies if the observer provides platforms with the predicted *probabilities* generated by the machine learning classifier rather than the predicted *classes*. We simulate the relative value that such an approach can provide to platforms with heterogeneous governance policies and find that it is superior to the value obtained from a range of alternative approaches.

A natural question that arises is how behavior of the actors may evolve, and specifically whether domain registrants who intend to produce illegitimate content (e.g., disinformation) can change their behavior to avoid detection by the external observer. We examine this question with a game theoretic model that leverages the differential cost of transparency between illegitimate and general domain registrants. We adapt a standard signaling game model to our setting and pare it down to its simplest form to facilitate intuition on the role of

Downloaded from informs.org by [193.60.238.99] on 07 April 2024, at 08:14 . For personal use only, all rights reserved.

transparency. The model allows us to assess the transparency levels in the domain registration information and how the observer, who has no direct control over the domain registration process, can exploit the differences to flag suspect domains. Intuitively, the results of the model highlight a tradeoff faced by illegitimate domain registrants between (1) revealing their domains as illegitimate domains and masking their own identities (a separating equilibrium), versus (2) masquerading their domains as general domains and revealing their own identities (a pooling equilibrium). We show that even when registrants have full information about the role of the external observer, the signals from the game theoretic model disadvantage illegitimate domain registrants, while general domain registrants are weakly better off.

Our research makes two contributions. First, our paper extends the literature on platform and ecosystem governance to consider a form of soft governance in low- or no-governance environments. Much of this literature focuses on governance that is constrained to the boundaries of the platform (Wareham et al. 2014, Zhang et al. 2019, Rietveld et al. 2021) that is largely managed by a central core actor (Boudreau and Hagiu 2009). However, many digital platforms also operate in, or allow exchanges with broader, ungoverned environments, such as the Internet. Rather than possessing a central core actor, the Internet consists of independent (often country-specific) agencies that set their own policies, and a variety of decentralized stakeholders—including those from academic, industry, and public sectors—that attempt to maintain shared standards and interoperability. We examine how an observer without formal authority can partially offset this regulatory void by assessing transparency signals from domain registration data. This is akin to establishing a reputation system, but one that is based on transparency. Goldfarb and Tucker (2019, p. 26) highlight the value of reputation systems by observing, “The rise of online reputation systems has facilitated trust and created new markets.”

Second, we complement recent studies that examine disinformation deterrence strategies. Research on measuring and modeling disinformation detection is typically in the context of a single social media platform (Vosoughi et al. 2018, Grinberg et al. 2019, Papanastasiou 2020) and largely focuses on interventions that target disinformation consumers through education (Guess et al. 2020), content tags (Moravec et al. 2020), fact-checks (Pennycook and Rand 2019), or content sourcing (Kim and Dennis 2019). Recent work on regulating online content acknowledges the possibility of governance regimes that do not reside solely on a platform (Cusumano et al. 2021). Our research formalizes this concept by proposing an external observer that can support multiple platforms to identify sources of disinformation before that content has infiltrated those platforms. Our empirical use of domain registration data are also systematically available and may offer a valuable proxy to researchers for the identification of disinformation (and other illegitimate) content.

2. Motivating the Soft Governance Observer

In this section we summarize the literature on platform governance and how it reflects layers within a broader governance architecture. We then identify how gaps in governance may emerge and how an independent observer can add value.

2.1. Overview of Platform Governance Literature

The literature on platform governance generally looks at questions of proper functioning within a platform or competitive implications across platforms. We consider three levels of prospective governance: platforms, governments, and regulatory intermediaries operating outside the platform in the broader environment.

Much of the research on platform regulation has focused on regulating the actions and outputs of actors within the platform to create and capture value for the platform (Rietveld and Schilling 2021). A platform is characterized by a “core” actor or manager that designs and implements rules governing the participating complementors and activities on the platform (Kretschmer et al. 2022). In regulating the platform’s complementors and their output, the core actor balances two countervailing forces (Ghazawneh and Henfridsson 2013, Wareham et al. 2014)—greater participation to facilitate innovation and variety (Boudreau 2012) versus greater control to establish standards and restrict low quality outputs (Boudreau and Hagiu 2009). Platform managers have a variety of tools available to manage this tradeoff, including recommending and promoting complementors (Fleder and Hosanagar 2014; Rietveld et al. 2019, 2021), establishing restrictions (Casadesus-Masanell and Halaburda 2014), mandating quality standards (Boudreau and Hagiu 2009), and encouraging desirable behaviors (Claussen et al. 2013). To further ensure standards and behaviors, the platform can apply verification systems to either the participant (Wang et al. 2021) or the underlying product or transaction (Kokkodis et al. 2022). Verification systems have traditionally been operationalized as ratings or reviews (Tadelis 2016, Goldfarb and Tucker 2019). Academic (Kim and Dennis 2019, Moravec et al. 2019, Pennycook et al. 2020) and policy proposals (Persily and Tucker 2021) attempt to curtail illegitimate content by tagging it at the point of entry onto the platform.

A second research stream focuses on governance that is administered and enforced by government agencies. This relates primarily to the platform’s competitive practices (see Jacobides and Lianos 2021, and cites therein), and considers market power, concentration, and winner-take-all dynamics that are driven by network effects (Bamberger and Lobel 2017, Song 2021). This line of

Downloaded from informs.org by [193.60.238.99] on 07 April 2024, at 08:14 . For personal use only, all rights reserved.

research is in the tradition of work in economics on monopolistic practices (Kovacic and Shapiro 2000). There are also legal regimes for verification, such as trademarks or other intellectual property, that offer participants in digital environments the opportunity to verify transaction partners (Bechtold and Tucker 2014).

A final research stream deals with regulatory intermediaries that may provide regulatory support for transactions between organizations operating in a broader environment. In an environment, this form of governance may formally confer specific responsibilities to an actor or the responsibilities may be assumed without a formal mandate. The literature on ecosystems conceives of an "ecosystem leader" who is a participant in the ecosystem and guides its norms and structures (Adner 2017). Governance in ecosystems can promote trust by effective management of partner interactions (Ruokolainen et al. 2011). In open-source communities, a balance between bureaucratic and democratic structures (O'Mahony and Ferraro 2007) facilitates the decision and property rights amongst participants (Shah 2006). Decentralized autonomous organizations (which are effectively confederations of actors) can manage governance with no central actor, and instead use algorithmic, social, and goal coordination (Hsieh and Vergne 2023).

2.2. The Governance Gap

Platforms and their environments are dynamic, and despite the multilayered nature of governance, gaps in governance coverage can emerge. One factor contributing to the introduction and dissemination of illegitimate content or transactions into platforms is exchanges among those platforms and their environment, which may host illegitimate actors. Illegitimate actors may establish a presence or content in the outside environment or on other platforms (Wilson and Starbird 2020) and use those footholds to infiltrate other platforms. Even an otherwise well-governed platform may be vulnerable to these cross-boundary incursions. The evolving cryptocurrency environment provides a specific example. Cryptocurrency adherents espouse an ethos of self- or no-regulation, which is reflected in its initial period of adoption and growth. The lack of oversight, however, allows illegitimate actors to establish low-quality tokens in the environment that can be used to defraud investors (Hamrick et al. 2018). When those tokens are traded against higher-quality tokens on higher-quality exchanges, it can undermine those platforms and ultimately the entire environment.

A second factor that can lead to illegitimate content is that platforms do not have jurisdiction over content that is outside of the platform, and governments may lack interest (for example, the initial lack of regulation of cryptocurrency exchanges in the United States) or authority to regulate some types of content (for example, the difficulty in constraining disinformation due to free speech considerations; U.S. Court of Appeals for the Fifth Circuit 2023). By hosting illegitimate content outside of platforms, and using platforms solely to promote that content, illegitimate actors are exploiting gaps in the governance structure.

2.3. The Soft Governance Observer

We propose an independent and external observer that enhances the digital environment's governance through transparency. Platforms and ecosystems assume that there is some form of "alignment structure" among actors (Adner 2017), but this concept may not account for actors who introduce illegitimate content or transactions in a broader environment in order to manipulate other participants and platforms. Those actors insulate themselves from sanction by creating this content using proxies, such as shell companies, websites, or applications. This can be mitigated by an observer who serves as a regulatory proxy by assessing the sources of content in the environment.

The observer is unlikely to be endowed with any formal authority in the environment. As a consequence, the observer may not control access to the environment or the activities of the actors, nor censor or alter content within or across platforms. Instead, the observer focuses on monitoring and assessing content *sources* before content is generated or enters a platform. This upstream focus complements downstream efforts by platforms and third parties to tag content after it has been created (Moravec et al. 2020). The observer achieves its upstream objective by using existing data about the sources of content in the environment, such as registration information for each source, to assess whether sources can be attributed to actors. An actor may choose to facilitate or avoid public attribution when establishing a source of content in an environment. It is possible that legitimate actors may have valid reasons for avoiding attribution, such as a dissident who prefers anonymity when sharing content in protest of an authoritarian regime. Such circumstances can be addressed by the observer using a transparency mechanism that allows actors to privately facilitate attribution of sources to those actors. The observer could eventually provide such a verification mechanism more broadly. An example of this in the nondigital world is the PreCheck® service offered by the Transportation Security Administration (TSA). In this case, people willingly provide verifiable information (and even pay for the privilege) to the TSA in exchange for expedited clearance when boarding flights in the United States. Similarly, the observer could offer a certification to positively influence a domain registrant's participation in more active verification processes.

2.3.1. Advantages. Separation and independence from individual platforms and other authorities might provide advantages to the observer that could improve outcomes on individual platforms and overall social welfare. First,

Downloaded from informs.org by [193.60.238.99] on 07 April 2024, at 08:14 . For personal use only, all rights reserved.

the observer's attribution efforts are platform-agnostic and portable across multiple platforms, which benefits legitimate actors whose content enters multiple platforms. Such multihoming actors would not need to obtain and maintain attribution for their content on platforms individually.

Second, the observer's activities could extend existing verification systems that exist within platforms. For example, by coordinating across platforms, the observer may also be able to leverage information and data from multiple sources that can yield greater value than any single source. The coordination would also allow for legitimate actors to obtain the aforementioned portable verification or accreditation that would reduce frictions or transaction costs. Thus, verification by an external observer passed into a platform could counteract internal reputation failures, such as ratings manipulation (Mayzlin et al. 2014).

Third, platforms are economic actors with their own set of value creation and capture incentives, which may or may not align with the mitigation of malicious content. If the platform is already monitoring its external environment (Nimmo and Torrey 2022), the observer could supplement the oversight effort taking place within the platform. Since the observer can operate as a shared service, it could enable greater economies of scale or streamline interactions with those being observed, including standard setting and resolving false predictions. The observer's independence also allows it to promote consistency and best-practices sharing across heterogeneous regimes. This can reduce the platform's compliance costs and provide a benchmark for outside parties to infer the internal policies of different platforms. By comparing actors and content that are (and are not) allowed on different platforms, relative to the recommendations of the observer, a researcher could better establish a platform's policy positions, or evaluate whether platforms are following through on monitoring efforts (Ricart et al. 2020). Observing changes to these policies over time may also provide insight on whether and how a platform's priorities change.[1] This type of indirect inference of internal policies offers an alternative to proposals that call for internal access of platform algorithms and policies by outsiders, such as researchers (Persily and Tucker 2021). Following the external observer's guidelines may also improve trust among users on that platform (Gu and Zhu 2021).

2.3.2. Temporal Dynamics. The observer must manage its activities amidst temporal dynamics. In particular, regulatory regimes may change and illegitimate actors may seek to erode the observer's effectiveness in unanticipated ways. To counter these eventualities, the observer must update its monitoring processes over time. This could be accomplished by implementing a feedback loop from the consumers of the observer's assessments, periodically updating and retraining the observer's assessment models, and refining or expanding the assessment escalation processes to account for changes in the environment. The observers can collect such feedback from individual platforms or its own assessment of the performance of its prior predictions.

3. Empirical and Analytical Approach to Understanding the External Observer

To gain some insight into the practical implementation of the proposed observer, we look at two separate, but related aspects of its operations. First, we confirm that an observer can effectively execute its monitoring function. We achieve this by building a machine learning classifier that distinguishes between legitimate and illegitimate content providers using innocuous data that is systematically available. We then use a simulation to reflect how platforms with different governance regimes could employ the classifier predictions. This allows us to assess how the observer might serve multiple platforms and confirm that each benefits from the observer's activities. Second, we look at how actors might respond to the presence of the observer. We show this using a game theoretic signaling model that captures the dynamic interactions between the observer and the actors operating in the environment. The signaling model confirms that despite having complete information about the intentions of the observer, illegitimate actors suffer from the observer's actions while legitimate actors weakly benefit from them.

3.1. Testing Transparency Using a Machine Learning Classifier

We examine these issues in the context of online disinformation. Disinformation has proliferated in the digital environment, and its presence erodes trust in organizations, the broader environment, and civil society (Waldrop 2017, Grinberg et al. 2019). However, countering disinformation on social media platforms is challenging. Such content can originate outside of platforms—most notably on websites—and circulate on the platforms through multiple automated or human accounts. The May 2021 Facebook (Meta) Threat Report on Influence Operations identifies platform diversification and the use of websites as emerging and important tactics of influence operation actors, stating "to evade detection and diversify risks, [influence] operations target multiple platforms (including smaller services) and the media, and rely on their own websites to carry on the campaign even when other parts of that campaign are shut down by any one company" (Gleicher et al. 2021b, p. 5). *The New York Times* reports that this trend toward deploying disinformation on websites is fed by the realization that disinformation articles and the websites that host them are more difficult to target, and thus, combat (Rosenberg and Barnes 2020).

The Internet Corporation for Assigned Names and Numbers (ICANN), which is responsible for policy and

Downloaded from informs.org by [193.60.238.99] on 07 April 2024, at 08:14 . For personal use only, all rights reserved.

technical management of the Internet's Domain Name System (DNS), is arguably the closest thing to a regulatory intermediary for the Internet. ICANN asserts, however, that "internet governance should mimic the structure of the Internet itself—borderless and open to all" (ICANN 2013) and "ICANN does not control content on the Internet. It cannot stop spam and it does not deal with access to the Internet" (ICANN 2021). This decision to forgo content controls supports ICANN's stated objective for Internet openness and creates an environment where individual platforms must apply their own governance policies to regulate content that originates from domains. However, it also allows for unchecked proliferation of illegitimate content in the environment. Illegitimate content on the Internet can include content that is malicious or wrongful, such as illicit transactions (e.g., phishing), malicious attacks (e.g., denial of service attacks), or disinformation. Such content can originate from actors who may have a presence on one or more platforms (Bakos and Halaburda 2020) or on nonplatform environments. The content can then be seeded onto well-governed platforms by the actor, its surrogates, or unwitting third parties.

The scope of this challenge is daunting. Over 200,000 domains are registered every day. Assessing the content of each domain would require a massive effort. However, a mechanism that easily sorts these domains based on the domain's probability of eventually hosting illegitimate content could lighten the burden. To achieve this, we distinguish between illegitimate domains and the illegitimate actors (the domain registrants) that establish the domains.[2] We propose using registration data provided by domain registrants to assess which domains have a disproportionate likelihood of hosting illegitimate content in the future. This early warning allows outside parties, platforms, and platform stakeholders to determine whether to subject the content from those domains to a higher level of scrutiny, such as algorithmic or human evaluation of their early content and traffic, or a network analysis of their early links to assess whether they are associated with other confirmed or suspected disinformation domains.

3.1.1. Classifier Design. We first establish the technical feasibility of the observer's function by developing a machine learning classifier that can help to disambiguate disinformation domains from legitimate domains. We assemble a data set of known disinformation domains and a random sample of general domains. Our sample of disinformation domains comes from Allcott and Gentzkow (2017). They collected links to articles that were proven to be false by Snopes, PolitiFact, or BuzzFeed, around the time of the 2016 U.S. presidential election and 2018 midterm election. While not a random sample, Allcott and Gentzkow (2017, p. 219) describe their database on disinformation as "a reasonable but probably not comprehensive sample of the major fake news stories … " The occurrence of disinformation campaigns around the 2016 U.S. presidential election was widely covered in the media (Timberg 2016, Shane 2017) and governmental reports (U.S. Office of the Director of National Intelligence 2017). The database consists of articles posted on 375 distinct domains. We drop 13 domains that hosted a single article that was shown to be false by the fact checking services, but otherwise provide credible news. The dropped domains are: bloomberg.com, dailymail.co.uk, huffingtonpost.com, huffingtonpost.co.uk, independent.co.uk, nydailynews.com, nymag.com, nypost.com, people.com, slate.com, talkingpointsmemo.com, washingtontimes.com, and buzzfeed.com. We drop eight more domains that were subdomains on aggregators or content creating platforms, namely Pocket, YouTube, WordPress, and BlogSpot. The final sample consists of 354 disinformation domains. The registration year for the disinformation domains in our sample are concentrated in 2016 and otherwise dispersed from 2006 through 2015 and 2017–2018.

We partnered with DomainTools, an online security and data company, to generate a sample of general domains for inclusion in our analysis. DomainTools generated a random sample of 75,000 domains whose registration periods roughly approximated the registration periods of the known disinformation domains. DomainTools randomly selected 30,000 domains from 2016 registrants and 45,000 domains from registrants in the other periods, matching the distribution of disinformation domains. From this set of 75,000 domains, we randomly sampled 4,000 domains for inclusion in our sample. We used the DomainTools "Who Is History" application programming interface (API) to download registration information on the 4,000 domains, from the first registration event of the domain. Among the 4,000 domains, complete information was available for 3,990 domains.

Our final sample consists of 354 disinformation and 3,990 general domains that we identify using a binary measure for disinformation ("1") or not ("0"). To address the imbalance between the classes, we use Synthetic Minority Over-Sampling Technique (SMOTE), a standard procedure in machine learning that over-samples the minority class by matching each observation to its nearest neighbors and under-samples the majority class (Chawla et al. 2002). We follow prior work and select over- and under-sampling percentages of 400% and 200% (Van Vlasselaer et al. 2017).

We used the DomainTools API to extract registration information for all disinformation and general domains in our sample. This includes the domain name, the extension, contact details provided by the registrant, the site, billing, and technical administrators, the date of registration, and the registrar. We performed the following feature engineering on the registration information. First, we use the domain name to compute the length of the

Downloaded from informs.org by [193.60.238.99] on 07 April 2024, at 08:14 . For personal use only, all rights reserved.

name, a count of the number of hyphens, a count of the numerical digits, an indicator for whether the number of periods exceeds one, and indicators for common domain extensions. We include squared terms and all two-way interactions of the continuous measures. A second set of features is based on the domain registrar. We categorize the registrar based on whether its frequency in the data are high (greater than 100), medium (between 10 and 100), or low (less than or equal to 10). A third set of features is based on the registrant's information. We categorize the registrant's name as being private, disclosed, or missing. We use the registrant's geographic information to create categories for each U.S. state (and "other" if the domain is international) and categories for each country (aggregating countries that only appear once and using "missing" if the country field is missing). We separately interact the registrar frequency features and the registrant name features with the U.S. state, country, domain extension, domain length, domain hyphens, and domain digits. We also interact the registrar and registrant features with one another. Our feature engineering exercise yields 1,139 features for the machine learning model. Of these, 252 are continuous, discrete, or binary measures, and 887 are interaction measures. Table 1 summarizes the measures obtained from the registration information.

A challenge with machine learning algorithms is that the resulting classifiers can fit the observed training data well but are not generalizable to new data in a production environment. We guard against this concern in two ways. First, we hold out a random sample of 20% of the observations for the test stage to estimate the performance of the classifiers. The other 80% of the data are used to train and validate the classifier using k-fold cross validation. Second, we use an elastic net penalized logistic regression algorithm to develop our classifiers (Zou and Hastie 2005). The algorithm is similar to the maximum likelihood estimator, but it imposes a compound penalty on the feature weights to mitigate overfitting the model to the training data. We confirm that our results are robust over a variety of different design choices, including splits in the training versus testing datasets, hyperparameter settings, and the empirical context.

The elastic net algorithm solves:

$$\hat{\beta} = \underset{\beta_0, \beta}{\operatorname{argmin}} \left[\frac{1}{N} \sum_{i=1}^{N} \ell_i(y_i, \beta_0 + x_i'\beta) \right] + \lambda \left[(1-\alpha) \sum_{j=1}^{J} \beta_j^2 + \alpha \sum_{j=1}^{J} |\beta_j| \right] \quad (1)$$

In Equation 1, $\ell_i(\cdot)$ is the negative log-likelihood for a logistic regression over N observations. The terms y_i and x_i are the outcome value and the set of feature values for observation i. The intercept and feature weights are represented by β_0 and β, respectively. The hyperparameters of the elastic net algorithm are λ and α. The first hyperparameter, $\lambda > 0$, controls the extent to which coefficients are penalized in the algorithm. The second hyperparameter, $0 \leq \alpha \leq 1$, governs the weight between a ridge and a Lasso penalty term. The ridge penalty is the sum of the squares of the feature weights (with β_j being the value of the weight on feature j) over J features, while the Lasso penalty is the sum of their absolute values. We set $\alpha = 0.99$, which is essentially a Lasso classifier without erratic behaviors that arise from highly correlated variables (Friedman et al. 2010).

Table 1. Summary of Domain Registrations Data

Panel A: summary of measures from registration data

		Distribution by frequency		
Variable	Unique fields	25% of data	50% of data	75% of data
U.S. state	49	1	1	2
Country	76	1	2	4
Domain extension	116	1	2	11
Registrar frequency	4	1	2	3
Registrant privacy	3	1	1	1

Panel B: measures computed from domain name (n = 4,344)

Feature	Mean	Std dev	Min	Max
Length	12.164	5.865	2	50
Multiple periods	0.080	0.271	0	1
Dashes	0.176	0.590	0	7
Digits	0.564	1.696	0	16

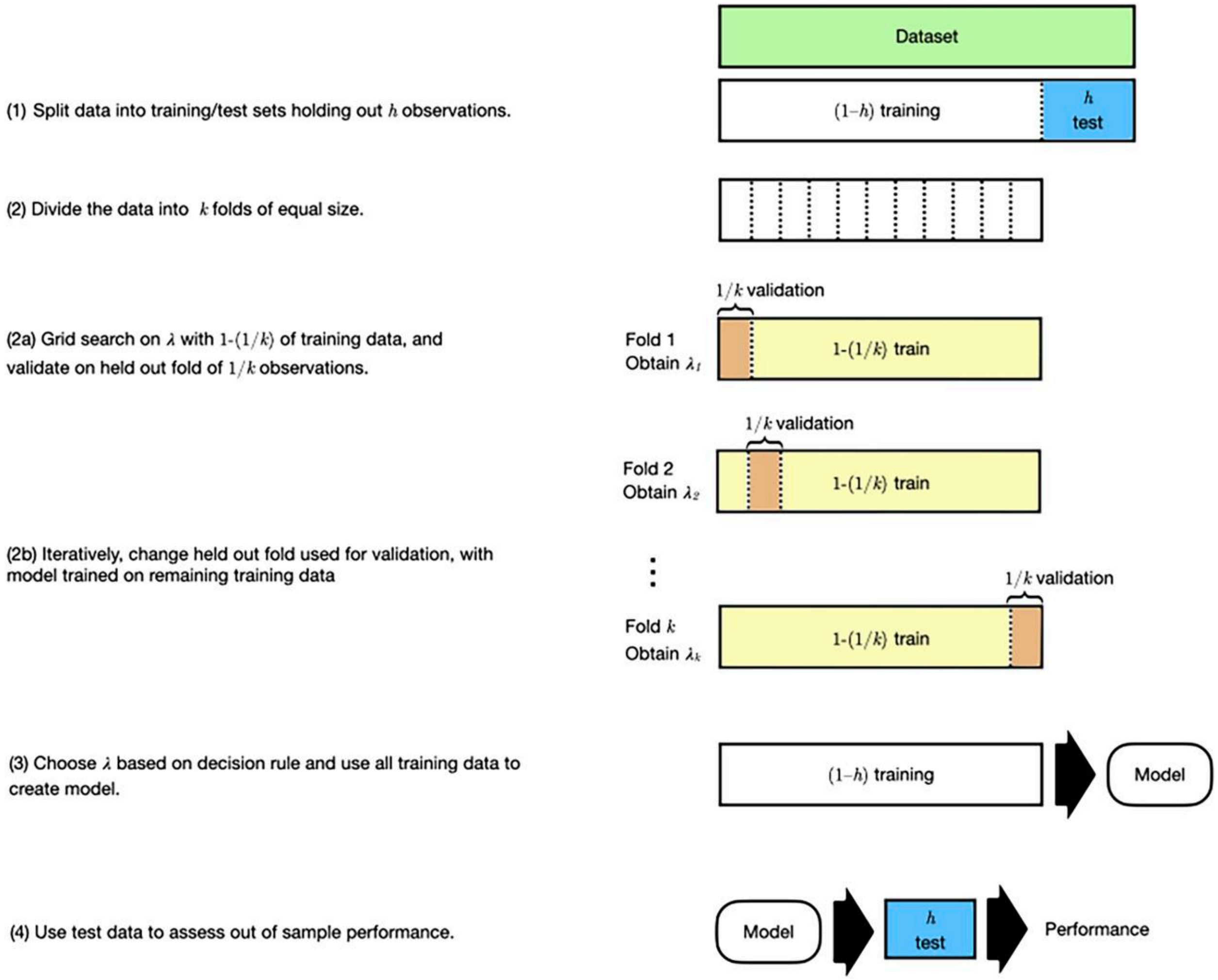
Notes. The distribution by frequency provides indicates how many unique values for each variable are identified by quartile. For instance, between 25% and 50% of all observations are from the same country, and between 50% and 75% of all observations are from the two most frequent countries.

Downloaded from informs.org by [193.60.238.99] on 07 April 2024, at 08:14 . For personal use only, all rights reserved.

We adopt standard processes in the machine learning literature for our model development (see Hastie et al. 2017, for details). Figure 1 summarizes the model development process. First, we separate the data into training and test datasets, holding out a proportion, h, of observations as the test data set. For our main results, we use $h = 0.20$. We then train an elastic net classifier using only the remaining $1 - h$ observations as the training data set. The training process includes k-fold cross-validation (kFCV, where we use $k = 10$) to select the hyperparameter, λ, in the elastic net classifier. For this cross-validation, the algorithm performs a grid search in each of the k iterations. For each iteration, the training data are divided into a train set, comprising $(k-1)/k$ of the observations, and a validation set, comprising $1/k$ of the observations. Across the k iterations, each observation appears in exactly one validation set and $k-1$ train sets. This yields k different classifiers corresponding to each of the train sets. The value of λ corresponds to that of the classifier with the fewest features (i.e., the most parsimonious classifier) whose average mean squared error (MSE) is within one standard error of the minimum MSE. This value of λ is then used to train the classifier using all of the training data. Finally, we use the resulting classifier against the hold-out test data to assess and report performance.

3.1.2. Results. The resulting elastic net classification model has nonzero weights on 193 features, including 144 features that are interactions of multiple characteristics. These include characteristics of the domain name, the registrant, and the registrar. The number and variety of features in the final classifier and the prevalence of interacted features underscore that the machine learning algorithm is identifying combinations that may be nonobvious in a human review of registration data. The features in our data are either binary or we have normalized them to a mean of 0 and a standard deviation of 1, so the magnitude of a feature's weight reflects the impact the feature has on the model's prediction. Over the 193

Figure 1. (Color online) Summary of k-Fold Cross-Validation Process

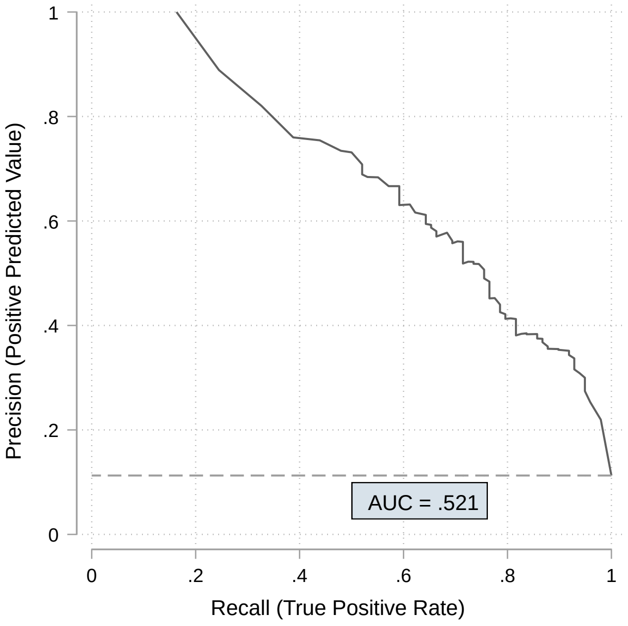

Dataset
(1) Split data into training/test sets holding out h observations.
(1–h) training
h
test
(2) Divide the data into k folds of equal size.
1/k validation
(2a) Grid search on λ with 1-(1/k) of training data, and validate on held out fold of 1/k observations.
Fold 1
Obtain λ1
1-(1/k) train
1/k validation
Fold 2
Obtain λ2
1-(1/k) train
(2b) Iteratively, change held out fold used for validation, with model trained on remaining training data
1/k validation
Fold k
Obtain λk
1-(1/k) train
(3) Choose λ based on decision rule and use all training data to create model.
(1–h) training
Model
(4) Use test data to assess out of sample performance.
Model
h
test
Performance

Downloaded from informs.org by [193.60.238.99] on 07 April 2024, at 08:14 . For personal use only, all rights reserved.

features in the model, the absolute values of the weights range from 0.001 to 7.080, with a mean 1.219 and a median 0.633. Features related to registrant identification information (i.e., features involving the registered name, contact details, address fields, etc.) are twice as likely to have nonzero weights in the classification model compared with other features and comprise 156 out of the 193 features with nonzero weights, including the fourteen features with the largest weights. This reinforces our proposition that identity transparency, or lack thereof, helps the classifier detect disinformation domains.

The classifier generates a predicted probability that a domain will produce disinformation in the future. By comparing the predicted probabilities to a cutoff threshold ($\theta \in [0, 1]$), we can generate a predicted class for each domain. If $\theta = 0$ is chosen, then all of the domains are classified as disinformation, which yields only true positive and false positive outcomes. If $\theta = 1$ is chosen, then none of the domains are classified as disinformation, which yields only true negative and false negative outcomes. A value of θ between 0 and 1 can result in a mix of true positive, false positive, true negative and false negative outcomes. For each threshold value of θ, we denote the proportion of all outcomes that are true positives as TP_θ, false positives as FP_θ, true negatives as TN_θ, and false negatives as FN_θ. In Figure 2, we present the precision-recall (PR) curve, which maps the classifier's precision ($TP_\theta/(TP_\theta + FP_\theta)$) and recall ($TP_\theta/(TP_\theta + FN_\theta)$) at values of θ from 0 to 1. The classifier's performance can also be depicted by a receiver operating characteristic (ROC) curve (available from the authors).

Figure 2. (Color online) Precision-Recall Curve for Elastic Net Classifier of Disinformation Domains



Note. Dashed lines represent performance of a random classifier.

However, PR curves are more conservative, and thus preferred, in instances where there is class imbalance because the PR curve is not dependent upon true predictions of the overweighted class (actual negatives, in our case). To assess the performance of the classifier along these dimensions, the AUC of our classifier's PR curve can be compared with AUC for the PR curve of a random classifier. The latter is equivalent to the proportion of disinformation domains in the holdout data. Our classifier generates a PR AUC of 0.521, which represents strong predictive power compared with the PR AUC of the random classifier of 0.112 (denoted by the dashed line in Figure 2). These results validate that domain registration information can be used to distinguish between disinformation and general domains at the time a domain is registered and before it starts to host content.

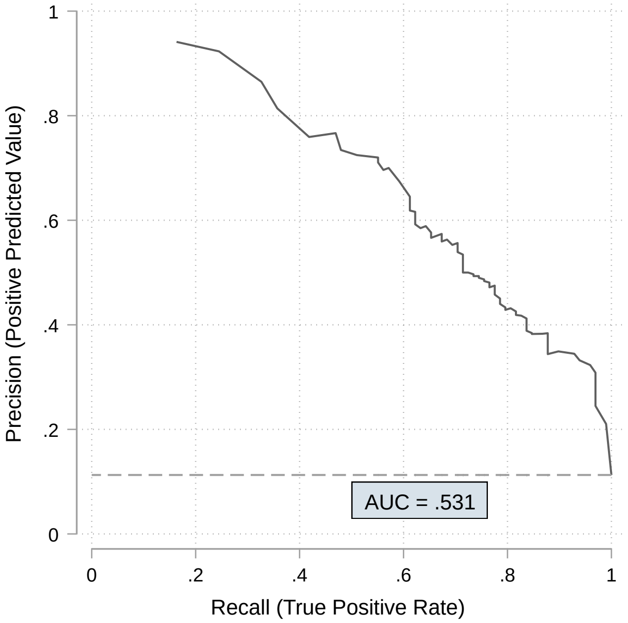
3.1.3. Results Using Identity-Related Features. To confirm that our classifier's performance is driven by the identification information included in domain registration data, we train a new classifier using only features that are related to the identity of the registrant. This represents a subset of the features described in Section 3.1.1 that are based on the registrant's contact details as well as any interaction terms that include this information. This reduces the set of features to 794. We follow the same process to train an elastic net algorithm and test the resulting classifier on the same holdout data set. The resulting elastic net classification model has nonzero weights on 185 features.

Figure 3 presents the PR curve using the classifier based on this constrained feature set. The shape of the PR curve is very similar to the PR curve for our main model (presented in Figure 2). The AUC for the PR curve is 0.531 (compared with 0.521 for our main model). Note that the predictive performance for this reduced-feature model slightly exceeds that from our main model. Generating a better predictive performance with less data may seem counterintuitive. This small performance difference reflects the data set shift (i.e., differences between the training and test datasets) that is almost universally present in practical settings. It underscores the importance of presenting predictive performance that is based on a holdout data set, as we have done. These observations provide strong evidence that the classifier's predictive performance is largely driven by the identity features that are extracted from the registration data.

3.1.4. Alternative Classifier Design Choices and Setting. We run a series of robustness checks to assess whether our classifier's predictive power is an artifact of our algorithmic design choices. We implement two types of variations: the proportion of data held out for testing and the value of the α hyperparameter used in the elastic net classifier. In our main results, we hold out 20% of the data in the sample in order to assess the performance of

Downloaded from informs.org by [193.60.238.99] on 07 April 2024, at 08:14 . For personal use only, all rights reserved.

Figure 3. (Color online) Precision-Recall Curve for Elastic Net Classifier of Disinformation Domains Using Only Identity-Related Features From the Domain Registration Data



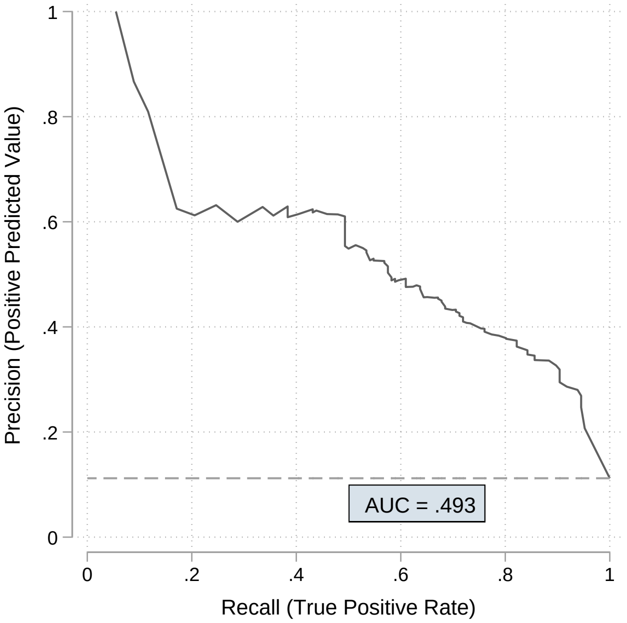
Note. Dashed lines represent performance of a random classifier.

the classifier trained on 80% of the data. Holding out more data would result in more observations on which the classifier's performance can be assessed, but less data on which the classifier can be trained. To check whether our model's performance is sensitive to the proportion of data held out for testing, we create a classifier using a holdout of 30%. Performance results are shown in Figure 4. The PR AUC of the model is 0.493.

Second, we evaluate alternative measures for the α hyperparameter, which controls the relative weighting between a Lasso and ridge penalty term in our elastic net classifier. Our main results are based on $\alpha = 0.99$. In practice, an analyst may use the data to determine the value of α, as part of the cross-validation process, or choose an alternative value. We evaluate the model's performance using $\alpha = 0$ (a ridge model), 0.25, 0.50, 0.75, and 1 (a Lasso model). The results are presented in Figure 5. For the resulting classifiers, the PR AUC varies from 0.502 to 0.615.

Another concern is that the results are limited to the context of disinformation domains. To assess whether other kinds of illegitimate actors can be identified by their registration behavior, we consider an alternative setting in which we predict the establishment of phishing domains. We perform a feature extraction and model building exercise, described in Online Appendix Section A, that is similar to the process we followed for disinformation domains. In this second setting we again find that registration information is informative in predicting whether the domain is set up for phishing activities.

Figure 4. (Color online) Precision-Recall Curve Using 30% Testing Holdout



Note. Dashed line represents baseline performance of a random classifier.

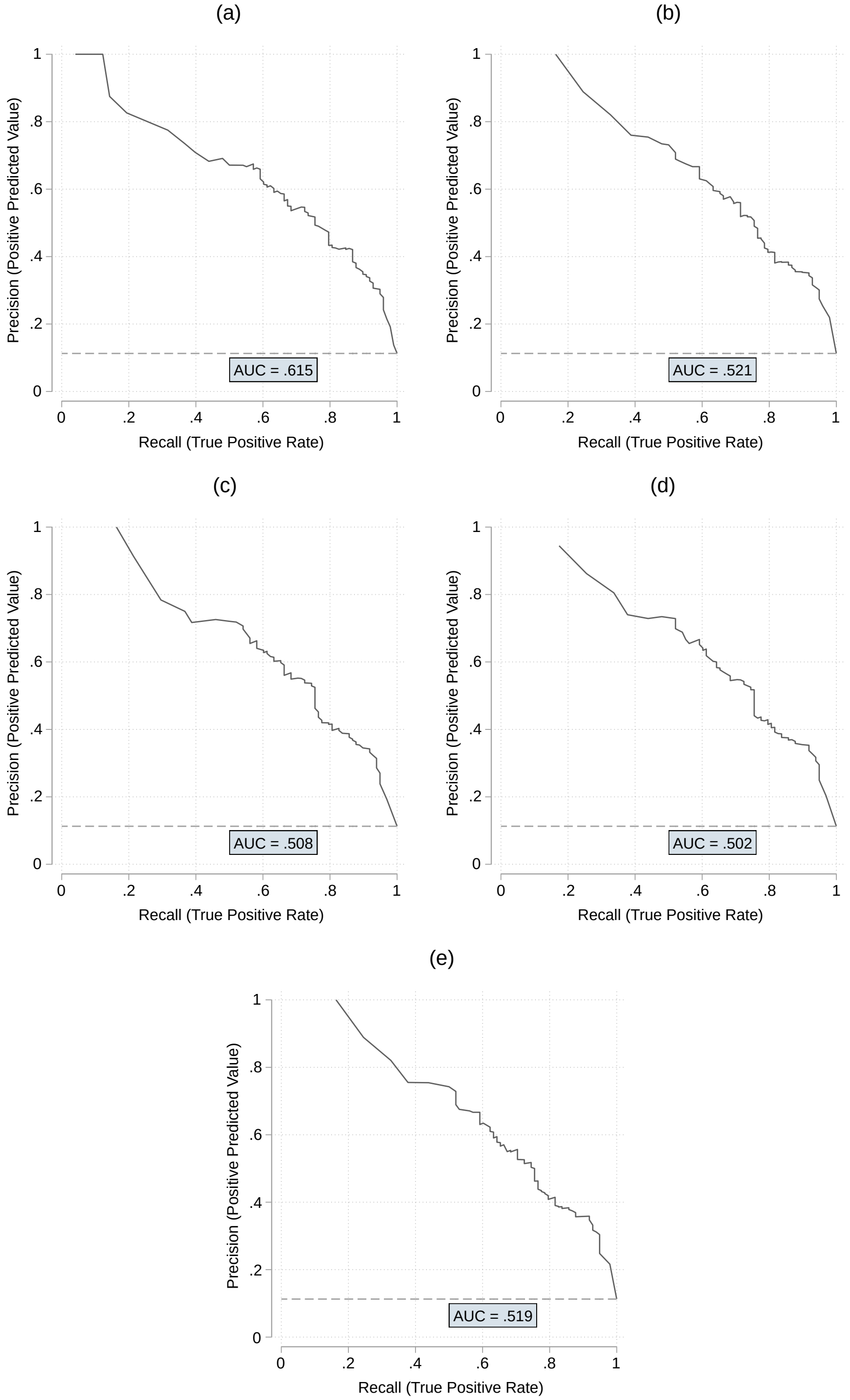
Collectively, these robustness checks confirm that the performance of our main classifier is not necessarily an artifact of our design decisions or empirical setting and that further model tuning may improve the classifier's predictive performance.

3.2. Optimizing Results for Different Governance Regimes

PR curves are effective at summarizing a model's predictive performance over alternative values of θ, but a single value for this threshold is used when a classifier is put into service. One possibility for choosing θ is that the observer selects the threshold by solving an optimization problem that balances social costs and benefits associated with a classifier's four prediction outcomes: true positives, false positives, true negatives, and false negatives. In this case, the benefit of true positives and true negatives, and the cost of false positives and false negatives, should reflect the governance priorities of all of the participants in the environment that the observer is assessing. However, applying a single economic model in support of individual organizations with different governance regimes may not yield locally optimal results. As Huber et al. (2017) observe, when implementing a governance regime, platforms must balance global rules versus adaptions that meet local requirements.

So how might the observer's assessment of disinformation domains be relevant to subscriber organizations with different governance regimes, such as different social media platforms, research organizations, security firms, policy makers, or government agencies? We demonstrate that this can be achieved, at least in part, by having those organizations subscribe to the observer's predicted probabilities rather than predicted classes for each domain. Doing so gives each subscriber the flexibility to apply its own tailored economic model to establish cutoff thresholds and determine each domain's predicted class. Organizations will have even more flexibility to tailor a predictive model to their needs if they simply in-source the observer function, but this may introduce other burdens, including operational and coordination costs.

Figure 5. (Color online) Receiver Operating Characteristic (ROC) Curves of Alternative Classifiers



Note. Dashed lines represent performance of a random classifier. (a) Ridge classifier. (b) Lasso classifier. (c) Elastic net classifier, $\alpha = 0.25$. (d) Elastic net classifier, $\alpha = 0.50$. (e) Elastic net classifier, $\alpha = 0.75$.

Downloaded from informs.org by [193.60.238.99] on 07 April 2024, at 08:14 . For personal use only, all rights reserved.

Downloaded from informs.org by [193.60.238.99] on 07 April 2024, at 08:14 . For personal use only, all rights reserved.

To examine the value of serving multiple subscribers by providing predicted probabilities, we consider an environment with one observer and multiple platforms with a *lax*, *moderate*, or *strict* disinformation governance regime. Each platform can select an optimal θ using an economic model of its utility function that aligns with its internal governance priorities and that captures the costs and benefits it incurs from the classifier's predictions. Equation 2 represents one example of how the economic model could be structured and serves as the basis for our simulations. Alternative model specifications can be easily accommodated. In this example, we reflect different governance regimes by varying the costs rather than the economic model specification. A platform incurs an expected benefit V, for identifying a disinformation domain (i.e., a true positive). Its policy is to put each predictive positive domain through additional verification at a cost C_v. Each false positive imposes a cost C_f on the platform, arising from frictions caused by exerting further scrutiny on an otherwise legitimate domain. The platform incurs a cost C_m, for misidentifying a disinformation domain (i.e., a false negative). We use TP_θ, FN_θ, and FP_θ to denote the true positive, false negative, and true negative proportions from the classifier's predictions, based on the value of θ. True negatives do not influence the economic model in this example, although it is trivial to include this and other adjustments if they more accurately reflect an organization's utility function. Combining these terms yields the utility from the classifier at different levels of θ, as follows:

$$\text{Utility}(\theta) = V(TP_\theta) - [C_v(TP_\theta + FP_\theta) + C_f(FP_\theta) + C_m(FN_\theta)]. \tag{2}$$

This equation serves as an objective function that the platform can apply against the information provided by the observer to determine the platform's optimal threshold. The platform can then classify newly or recently registered domains by applying its optimal threshold to the probabilities that the observer generates on new domains. This approach is akin to the H-measure approach of assessing classifier performance (Hand 2009) but includes the economic implications of true classifications as well.

We simulate a range of governance regimes by assigning different combinations of V, C_v, C_f, and C_m and using this information to solve Equation 2. In our simulations, we normalize $V = 1$ and allow C_f, C_v, and C_m to range from 0 to 2 in increments of 0.01. This yields a total of 8,120,601 simulations. Table 2 provides results using this process for three representative governance regimes, which we label *strict*, *moderate*, and *lax*. Additional details and a broader (but still partial) set of simulation results are in the Online Appendix. The full simulation results align with the presented examples and are available from the authors.

Table 2. Cost, Benefits, Confusion Matrices and Predictive Performance for Three Representative Subscriber Governance Types

Metric	Strict	Moderate	Lax
V	1.0	1.0	1.0
C_v	0.1	0.2	0.4
C_f	0.1	0.2	0.4
C_m	1.5	1.0	0.5
Optimal threshold	0.10	0.54	0.86
True positives	0.104	0.085	0.067
False positives	0.191	0.083	0.033
True negatives	0.696	0.804	0.854
False negatives	0.009	0.028	0.046
Sensitivity (Recall)	0.918	0.755	0.592
Specificity	0.785	0.907	0.962
Positive Predicted Value (Precision)	0.352	0.507	0.667
Negative Predicted Value	0.987	0.967	0.949
Accuracy	0.800	0.890	0.921
Incremental value (per 1,000 domains)			
Compared with naive options	117.6	119.8	47.3
Compared with Strict θ	—	9.0	85.7
Compared with Moderate θ	24.0	—	20.2
Compared with Lax θ	57.2	12.4	—

Notes. True positives, true negatives, false positives, and false negatives are presented as proportions of the predicted observations. Incremental values are a unitless measure estimated using *Utility*(θ) derived with Equation 2 based on the platform's optimal value of θ and subtracting *Utility*(θ) derived with Equation 2 based on the reference value of θ.

Table 2 presents four panes of information. The top pane contains the values of V, C_v, C_f, and C_m used for the simulation (presented as unitless numbers corresponding to the input values). The next pane identifies the optimal value of θ derived from Equation 2 and the resulting values in the confusion matrix. Intuitively, the optimal θ increases over the regimes, from strict (0.10), to moderate (0.54), to lax (0.86). The next pane provides common predictive performance measures—sensitivity (recall), specificity, positive predictive value (precision), negative predictive value, and accuracy. A higher value for each measure is desirable, but there is no simulation that has better performance across all the prediction measures.

The bottom pane presents the incremental value (presented as unitless numbers per 1,000 domains) that can be generated by using the optimal θ for the platform, relative to using other methods. We compute the incremental value by subtracting *Utility*(θ) using the platform's optimal value of θ from *Utility*(θ) based on (1) the highest result from among three naive prediction strategies: assume no disinformation domains (equivalent to $\theta = 1$), assume all disinformation domains (equivalent to $\theta = 0$), or randomly classified domains, (2) the optimal θ under a strict governance regime, (3) the optimal θ under a moderate governance regime, and (4) the optimal θ under a lax governance regime. By definition, *Utility*(θ) is maximized based on the platform's optimal value of θ, so it is not surprising that the values in this pane are all

Downloaded from informs.org by [193.60.238.99] on 07 April 2024, at 08:14 . For personal use only, all rights reserved.

positive. This pane is useful, however, in representing a relative scale for these results.

These results underscore the importance of tuning a machine learning classifier using a value measure that reflects the governance priorities of the subscribing organization, rather than tuning a machine learning classifier using predictive performance measures, such as accuracy. The results also make clear that if the observer intends to simultaneously serve multiple organizations (a centralized or outsourced service, for instance), then it can create more value by providing predicted probabilities to those organizations, rather than predicted classes based on an arbitrary cutoff threshold.

3.3. Game Theoretic Model

We develop and analyze a dynamic model of incomplete information to better understand the strategic behavior of registrants in these types of settings. The model provides an analytically tractable description of the external observer's role in establishing soft governance, potential changes in the behavior of domain registrants who intend to produce illegitimate content, and mechanisms that support more desirable outcomes. We largely conform to the canonical signaling game structure, as described in Connelly et al. (2011), and describe any modeling adjustments that we employ to reflect the idiosyncrasies our setting, including the belief refinement methods.

A critical feature of our model is to distinguish between the sources of illegitimate, deceptive content (Internet domains) and the actors (domain registrants) behind those domains. This modeling distinction aligns with other practical settings, such as the enforcement functions at social media platforms that disambiguate between deceptive content and operators exhibiting deceptive behavior. As Facebook's 2021 Threat Report points out "when a threat actor conceals their identity through deceptive behavior, the public will lack sufficient signals to judge who they are, how trustworthy their content is, or what their motivation might be" (Gleicher et al. 2021b, p. 3).

3.3.1. Model Design.

We model two players, a domain registrant (denoted R, she/her) and a risk neutral observer (denoted O, he/him). The registrant privately knows her type, $t \in \{L, H\}$. An H-type has high legitimacy and is interested in establishing a domain that is free of deception. An L-type has low legitimacy and is interested in establishing a domain that contains deception, for example, disinformation. It is common knowledge that the registrant is an H-type with probability $h \in (0, 1)$ and an L-type with probability $(1 - h)$. The registrant has some discretion on the composition of her registration information, that is, how she completes the registration, and the level of detail she provides. This aligns with the domain registration process. A registrant enters into a registration contract with one of approximately 2,000 authorized registrars and provides registration information that includes technical and contact information. ICANN requires that the registrant's contact information is accurate and reliable but leaves it to the registrars to confirm that this requirement is met. Identity verification measures are decentralized and susceptible to deception, providing ample room for registrants to obscure their identities. We use transparency $T \in \mathbb{R}^+$ to measure the degree to which the registrant's registration data provides verifiable information to identify the registrant's true identify. More (less) transparency makes it easier (harder) to attribute a domain to an individual or entity.

The observer evaluates domains at the time of registration but does not have control over the registration process and cannot prevent domains from operating. Upon assessing the registrant's transparency signal, the observer may update his belief that the registrant is an H-type with probability $\tau \in [0, 1]$ and an L-type with probability $1 - \tau$. Based on these updated beliefs, the observer's decision is to assign a monitoring level, scaled to $A \in [0, 1]$, that is applied to the registrant. The observer's utility function depends on the error of this assessment. Monitoring can take many forms in practice, including assigning a certification, subjecting the registrant's domain to increased scrutiny, or publishing a warning about the registrant's domain. Actions similar to a monitoring level can be seen in practice when social media companies label content as false or misleading, or Internet security companies publish lists of phishing domains. We capture the form of the observer's utility function following a simple structure for risk neutral actors in Gibbons (1992),

$$U_O(T, \tau) = -(A - (1 - \tau))^2. \tag{3}$$

This reflects the observer's desire to assign monitoring to the registrant that is commensurate with his perceived likelihood that the registrant will publish illegitimate content. The monitoring amount that maximizes the observer's utility is $A = 1 - \tau$. In equilibrium, the observer's updated belief resolves to three cases: the registrant is an H-type ($\tau = 1$), the registrant is an L-type ($\tau = 0$), or the observer maintains his prior belief ($\tau = h$).

We model the domain registrant's utility function using three components—the value of establishing a website, the value of providing transparency, and the impact of monitoring. This three-part structure allows us to disambiguate the relationship between the value to the registrant of providing transparency (a signal) and the cost imposed on the registrant by the observer's monitoring activity. The value of running a website, denoted $E(t)$, varies with the registrant's type and reflects revenue streams or nonmonetary benefits accrued to the domain registrant. The value of providing transparency, denoted

Downloaded from informs.org by [193.60.238.99] on 07 April 2024, at 08:14 . For personal use only, all rights reserved.

$P(t, T)$, varies with the registrant's type and the amount of transparency. In line with standard signaling game models, we assume that both the cost and marginal cost of providing transparency are lower for the H-type registrant compared with the L-type registrant (Mas-Colell et al. 1995). Since we represent P as a benefit, this implies that $P(H,T) > P(L,T)$ and $\partial P(H,T)/\partial T > \partial P(L,T)/\partial T \ \ \forall T$. The intuition for this modeling choice is that an H-type registrant can receive some positive operational value from providing transparency (for instance, by facilitating information flows with the registrar for billing, site maintenance, or support issues), although this value need not be monotonically increasing (for instance, due to privacy concerns). An L-type registrant, however, views transparency as less desirable since it can reveal her true identity and expose her to operational costs (for instance, sanction or legal action).

The impact of monitoring on the registrant's utility is commensurate with the amount of monitoring assigned by the observer. Since A weakly decreases in τ, we can represent the monitoring cost as a function of τ, $M(\tau)$.[3] The cost can even be negative at high levels of τ (for instance, a benefit derived from a "trusted domain" certification). Combining these terms, the registrant's utility function is expressed as:

$$U_R(t,T,\tau) = E(t) + P(t,T) - M(\tau). \tag{4}$$

Figure 6 summarizes the sequence of events. First, the registrant learns her type and then chooses the amount of transparency that she will provide through her platform registration information. The observer assesses the registrant's transparency, updates his beliefs, and applies a monitoring level to the registrant. Finally, both players realize their utilities.

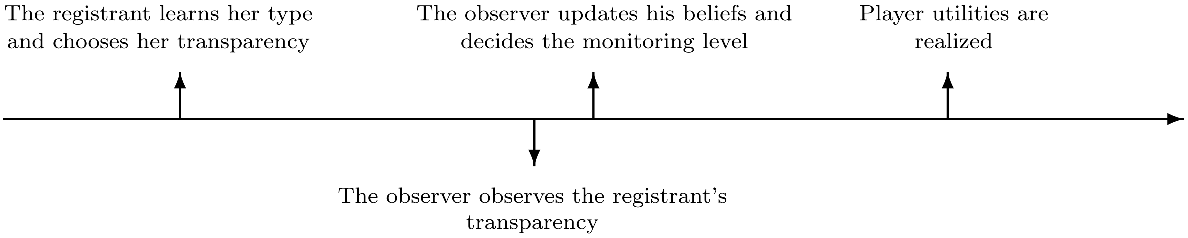
We utilize Perfect Bayesian Nash equilibrium (PBE) (Fudenberg and Tirole 1991) to define the players' strategies. A PBE requires that posterior beliefs adhere to Bayes rule. This may yield multiple equilibria, however. We pare down the list of unreasonable equilibrium using the undefeated refinement and lexicographically maximum sequential equilibrium (LMSE) (Mailath et al. 1993). This combination yields only PBE that (1) weakly improve the utilities for both types and (2) first prioritize the H-type registrant's preferred signal before prioritizing the L-type registrant's preferred signal. The intuition for this modeling choice aligns with our practical setting in which rational players are likely to gravitate toward pareto optimal outcomes and the L-type registrant wishes to masquerade as the H-type registrant (and will therefore mimic the H-type's behavior), rather than the opposite. In addition to providing reasonable and intuitive results, an LMSE yields a unique prediction in our setting. Our model aligns with the experimental evidence in Schmidt and Buell (2017), who show that decision makers more often select pareto dominant pooling equilibria over pareto dominated separating equilibria. Imposing alternative belief refinements, such as the intuitive criterion (Cho and Kreps 1987), does not alter our primary finding that transparency signals can be used to identify illegitimate content domains. However, the resulting separating equilibria may be pareto dominated by other equilibria that are eliminated by the intuitive criterion.

3.3.2. Outcomes.

We relegate the technical details of our analysis to the Online Appendix and focus here on the intuition for our results using a representative example. We compare the results of the signaling game to a reference case in which registration data are not employed as a means to distinguish illegitimate content domains from general domains. In the Section 3.3.3, we describe how the model can be extended to yield additional insights.

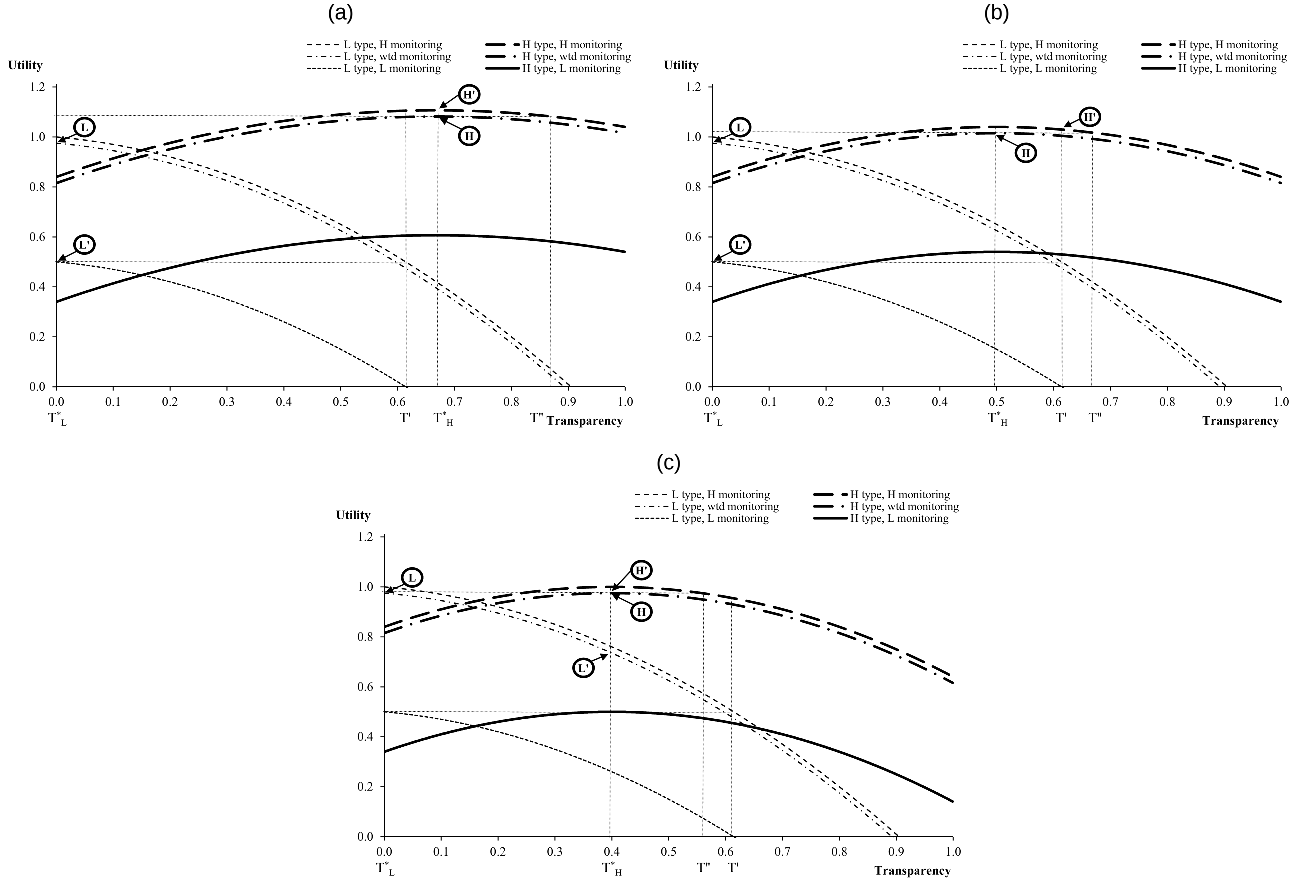
Figure 7 presents the transparency choices and utilities for the H- and L-types in the reference case compared with those in the signaling game. The x-axis represents the registrant's choice of transparency, the y-axis represents the registrant's utility. The curved lines map the registrant's utility function along these axes depending on the registrant's type and the observer's beliefs. Each subfigure presents six utility curves, three for the H-type registrant (thicker lines) and three for the L-type registrant (thinner lines). The three utility curves for each registrant type correspond to the observer's three possible equilibrium beliefs, that is, that the registrant is an H-type (the top dashed thick line and the top dashed thin line), is an L-type (the bottom solid thick line and the bottom dashed thin line), or the observer's prior belief that the registrant is an H-type with probability $\tau \in [0,1]$ and

Figure 6. Timeline of the Signaling Game


The registrant learns her type and chooses her transparency
The observer updates his beliefs and decides the monitoring level
Player utilities are realized
The observer observes the registrant's transparency

Downloaded from informs.org by [193.60.238.99] on 07 April 2024, at 08:14 . For personal use only, all rights reserved.

Figure 7. Representative Equilibrium Outcomes



Notes. H and L identify the transparency choices and utilities in the reference case for the *H*-type and *L*-type registrants, respectively. H′ and L′ identify the transparency choices and utilities in the signaling setting under a nondistortive separating equilibrium (Panel a), a distortive separating equilibrium (Panel b), and a pooling equilibrium (Panel c). We use $E(L) = 1$, $E(H) = 0.84$, $M(\tau) = (1 - \tau)0.5$, $P(L, T) = -0.2T - T^2$, and $h = 0.95$. The only difference is that $P(H, T) = 0.8T - 0.6T^2$ in Panel a, $P(H, T) = 0.8T - 0.8T^2$ in Panel b, and $P(H, T) = 0.8T - T^2$ in Panel c. (a) Nondistortive separating. (b) Distortive separating. (c) Pooling.

an L-type with probability $(1 - \tau)$ (the middle dashed thick line and the middle dashed thin line).

For illustrative purposes, we generate the utility curves using $E(L) = 1$, $E(H) = 0.84$, $M(\tau) = (1 - \tau)0.5$, $P(L, T) = -0.2T - T^2$, and $h = 0.95$. The only difference across the subfigures is that we modify the value of providing transparency for the H-type registrant, so $P(H, T) = 0.8T - 0.6T^2$ in Figure 7(a), $P(H, T) = 0.8T - 0.8T^2$ in Figure 7(b), and $P(H, T) = 0.8T - T^2$ in Figure 7(c). This adjustment monotonically reduces the value of providing transparency for the H-type and allows us to demonstrate each of the three possible forms of equilibrium outcomes—a separating PBE, a distortive separating PBE, and a pooling PBE.

Each subfigure identifies four transparency thresholds on the x-axis. These represent transparency levels that demarcate player behavior, making them important for understanding the equilibrium outcomes in the signaling game. The four transparency thresholds are (1) the L-type registrant's utility maximizing transparency level (T_L^*), given there is no information asymmetry about her type (i.e., $\tau = 0$), (2) the H-type registrant's utility maximizing transparency level (T_H^*), given there is no information asymmetry about her type (i.e., $\tau = 1$), (3) a transparency threshold (T') beyond which a low type's utility under a belief that she is a high type ($\tau = 1$) is dominated by her utility at T_L^* under a belief that she is a low type ($\tau = 0$), and (4) a transparency threshold (T'') beyond which a high type's utility under a belief that she is a high type ($\tau = 1$) is dominated by her utility at T_H^* under a weighted belief ($\tau = h$). Note that the ordering of these transparency thresholds along the x-axis can change depending on the model parameters, as evident in Figure 7. The technical details formally defining these thresholds are in the Online Appendix.

In all three subfigures, the equilibrium outcomes from the reference case (i.e., without signaling) are identified with an **L** for the L-type registrant and an **H** for the H-type registrant. Without signaling, the observer cannot identify illegitimate domains at the time of registration, so he is unsure of the domain's type and his posterior belief is equal to his prior belief, $\tau = h$. As a result, the

Downloaded from informs.org by [193.60.238.99] on 07 April 2024, at 08:14 . For personal use only, all rights reserved.

middle dashed thick line and the middle dashed thin line represent the utilities of the registrant types. The L-type registrant maximizes her utility by choosing T_L^* and the H-type registrant maximizes her utility by choosing T_H^*.

For the signaling game, each subfigure captures one of the three possible equilibria. Those outcomes are identified with an **L′** for the L-type registrant and an **H′** for the H-type registrant. Figure 7(a) represents a nondistortive separating equilibrium. The L-type continues to choose a transparency level T_L^*, and the observer recognizes that she is an L-type, so her utility falls to the bottom dashed thin line. The H-type continues to choose a transparency level T_H^*, and the observer recognizes that she is an H-type, so her utility increases to the top dashed thick line. In this equilibrium compared with the reference case, the L-type is worse off, the H-type is better off, and the uncertainty of the registrant's type is alleviated. This occurs because the value of providing transparency is positive for the H-type and costly for the L-type, so the H-type can identify herself by simply choosing her optimal level of transparency. The L-type has no recourse because it is too costly to mimic the H-type's transparency level.

Figure 7(b) represents a distortive separating equilibrium. The L-type continues to choose a transparency level T_L^*, and the observer recognizes that she is an L-type, so her utility falls to the bottom dashed thin line. The H-type chooses a transparency level $T' > T_H^*$, and the observer recognizes that she is an H-type, so her utility increases to the top dashed thick line. Once again, the L-type is worse off, the H-type is better off, and the uncertainty of the registrant's type is alleviated. However, the H-type decides to distort her transparency level above T_H^* to achieve this outcome. This occurs because the value of providing transparency has degraded somewhat for the H-type, which raises the plausible threat that the L-type could mimic her. To discourage the L-type from doing so, the H-type must provide this extra transparency, which benefits her through a lower monitoring cost compared with the reference case.

Figure 7(c) represents a pooling equilibrium. Both the L-type and the H-type choose a transparency level T_H^*, and now the observer cannot distinguish which type they are. As a result, the L-type's utility is on the middle dashed thin line, and the H-type's utility is on the middle dashed thick line. Compared with the reference case, the L-type is again worse off, but the H-type is no better or worse off, and the uncertainty of the registrant's type persists. This occurs because the value of providing transparency has degraded further for the H-type, such that the L-type can credibly mimic the H-type by investing in a transparency level beyond T''. At that point, the H-type is better off by reverting back to T_H^* and allowing the L-type to mimic her at that level.

In summary, comparing all three possible equilibrium outcomes to the base case, the L-type is worse off and the H-type is never worse off and can even be better off. The three equilibrium outcomes are representative of the model results generally and are not an artifact of this particular example.

3.3.3. Implications for the External Observer.

Extending our model can inform how the observer can further constrain illegitimate actors or introduce new countermeasures and predictive features. As a concrete example, we assume the maximum level of transparency is infinite ($T \in \mathbb{R}^+$). In practice, however, registration processes often implicitly limit the amount of transparency that can be provided due to structural choices in the design of the registration process. For instance, a registration process may require only simple contact information without verification. Such limits may constrain the ability of some legitimate registrants from voluntarily differentiating themselves from illegitimate registrants, leading to a pooling outcome. To see this in the model, consider the result if the maximum value of T is below T' and $T' < T''$, such that separating is desirable for the H-type, but impossible to achieve. Accounting for transparency constraints can inform how an observer can deliver more value to legitimate domains in exchange for participating in a more robust process to authenticate registration information. Allowing registrants to provide greater transparency can yield a nondistortive or distortive separating equilibrium, which can provide a higher utility for and H-type compared with a pooling equilibrium.

4. Discussion

We outline the environment in which platforms operate as consisting of a governance structure with three primary levels. The first level is provided by the platform itself and is established to support the platform's value creation function. The second level is provided by the government(s) which maintain jurisdiction over the platform. The third level is provided by regulatory intermediaries in the environment(s) in which the platform operates. Despite these multiple levels, governance gaps exist that can be exploited by illegitimate actors to introduce illegitimate content or transactions. We argue that an independent observer can mitigate this issue by collecting and assessing transparency signals from actors as they establish a presence in the environment. This assessment allows the observer to exercise a form of soft governance in which the observer exerts indirect influence without formal authority.

We examine this issue in the context of an observer whose function is to predict domains that will host disinformation. We argue that a disinformation registrant's desire to dissemble their true identities will manifest in

Downloaded from informs.org by [193.60.238.99] on 07 April 2024, at 08:14 . For personal use only, all rights reserved.

registration data and can signal a disinformation domain. Anecdotal evidence in related contexts support that disinformation providers take steps to disguise their identities. For instance, Meta's December 2021 Adversarial Threat Report confirms that identity concealment on Facebook is common with influence operations and documents several instances in which the entities behind recent coordinated disinformation campaigns tried to hide their identities on the platform (Gleicher et al. 2021a). Our approach exploits the information that domain owners already provide through existing domain registration processes and uses it as a signal of the domain's legitimacy. This is not a comprehensive solution to the problem of disinformation, nor is it immune to the concern that some registrants may wish to conceal their true identities for reasons other than being linked to disinformation. Ultimately, our research suggests that an observer that exploits voluntary transparency signals can raise the costs of disinformation providers, while simultaneously reducing the costs of legitimate information providers.

It is notable that registrants of some established disinformation domains are transparent about their identities. An example of this is Alex Jones, the registrant behind the disinformation domain InfoWars.com. In our signaling model, such a situation corresponds to a pooling equilibrium in which the registrant is transparent about their identity and whose domain may avoid being flagged by the observer at the time of registration. These domains will only be identified as providing disinformation after they have spread sufficient disinformation to attract regulatory scrutiny. This is an unfortunate outcome, and other registrants may seek to replicate it by providing transparent registration information. Our model, however, indicates that such transparency exposes the registrant to operational costs such as sanctions or legal action. This is the outcome for Alex Jones, who was found guilty in November 2021 of defamation for the disinformation that he spread about the Sandy Hook Elementary School massacre (Williamson 2021).

Training machine learning algorithms with registration information has already proven to be effective in other settings. Machine learning models are employed by social media platforms to identify illegitimate accounts at or near the time of registration (Bray 2018, Breuer et al. 2020). In those applications, however, the platform has complete control over its own registration process and can actively screen suspect registrants by imposing additional registration steps or collecting nonregistration data. In contrast, our approach addresses a shared environment that individual platforms cannot directly regulate and a registration process that they cannot control. Our approach offers hope that progress can be made in this environment. Meta's 2021 Threat Report serves as a call for coordinated action by stating, "We know that influence operations are rarely confined to one medium. While each service only has visibility into activity on its own platform, all of us—including independent researchers, law enforcement and journalists—can connect the dots to better counter IO [influence operations]" (Gleicher et al. 2021b, p. 5). We show how the independent observer's predicted probabilities can allow platforms to identify disinformation prior to its entry into their platforms and before it builds an audience.

Future work can expand beyond the limitations of our empirical study. For example, although our sample offers us the opportunity to test the efficacy of a machine learning model using relevant real-world data, it is not random. We believe the sample is appropriate to demonstrate the proof of concept. In practice, a model might initially be trained on nonrandom data and adapted with periodic retraining on new data. Future work may consider how to augment our practical findings in a hypothetical world with random assignment, perhaps by using an experimental setting in which subjects are randomly assigned different roles.

Our study does not eliminate the necessary evolution in the observer's operations as illegitimate actors seek to defeat its predictions, often referred to as adversarial responses in the machine learning literature. Such responses may result in data set shift, which represents changes in the distributions of the outcome or independent variables over time (Quiñonero-Candela et al. 2008). There is no universal cure to data set shift, but the literature points to several mitigation steps. These include retraining (Huang et al. 2011) and avoiding machine learning algorithms that may be particularly susceptible to such responses (Papernot et al. 2016). Future work can exploit the ongoing advances in the machine learning literature on this phenomenon. While our model was trained at a single point in time, an operationalized model should be retrained over time with newer data, and future training samples may be expanded to include other sources of disinformation. This can add greater resiliency to the model and may allow it to be tuned to identify particular types of disinformation.

While we present our results in the context of platform governance, the impact of allowing illegitimate actors to proliferate in these environments has consequences for everyone. The prevalence of disinformation is a phenomenon that individuals and firms will have to account for when managing their identity, reputation, and operations in digital environments. Instances of disinformation campaigns that target corporate and governmental organizations are now common and can originate from a variety of sources. For example, the U.S. Department of Homeland Security has highlighted domestic groups that organize disinformation campaigns to "do economic harm to a corporation with whom they disagreed" (U.S. Department of Homeland Security 2019). Facebook now regularly reports on its interdiction efforts against influence operations from multiple groups, including shutting down an operation in which one of south-east Asia's

Downloaded from informs.org by [193.60.238.99] on 07 April 2024, at 08:14 . For personal use only, all rights reserved.

largest telecommunication firms used Facebook accounts to conduct a commercial disinformation campaign seeking to discredit its competitors (Murphy and Reed 2020). And the U.S. government has accused other countries of engaging in far-reaching disinformation campaigns on the efficacy of commercial vaccinations for the COVID-19 pandemic (Barnes et al. 2020).

As more facets of organizations, digitization, and technology intersect, managers and scholars must contend with disinformation in the digital information environment. There are more opportunities for research in this area, including greater examination of how firms can restructure both technical and business processes to contend with the threat or presence of disinformation, how disinformation affects an organization's relationship with and management of its stakeholders, how to structure effective and coordinated governance measures within platforms and across environments, the competitive implications of espousing different levels of leniency toward disinformation and how different levels of leniency affect users and other platform participants. Considering disinformation could expand on related work in reputation formation (Cennamo and Santalo 2019, Etter et al. 2019), organizational relationships with stakeholders (Karunakaran et al. 2022), and platform performance and competition (Rietveld et al. 2021). The diversity of sources, topics, and targets for disinformation underscores the need for more boundary spanning solutions and research that can extend our understanding of platform and environment dynamics.

Acknowledgments

The authors thank seminar participants at Cornell University, the University of Michigan, and London Business School. The authors also thank Hunt Allcott and Matthew Gentzkow for sharing disinformation data, and DomainTools for sharing domain data. The authors thank attendees of the Wharton Tech Conference, Academy of Management, Strategic Management Society, and the Munich Summer Institute for their feedback.

Endnotes

[1] Recent news suggests that turnover in a platform's leadership may impact the platform's priorities as well as the priorities of other platforms (Nix and Ellison 2023).

[2] This distinction is implicitly acknowledged in a September 2022 announcement by Meta, the parent company of Facebook, that it had taken accounts off its platform that disseminated content from a "network of over 60 websites carefully impersonating legitimate websites of news organizations in Europe" (Nimmo and Torrey 2022).

[3] This cost can also be allowed to vary with the registrant's type, which reflects that a well-designed monitoring process can impose greater costs on an *L*-type registrant compared with an *H*-type registrant.

References

Adner R (2017) Ecosystem as structure: An actionable construct for strategy. *J. Management* 43(1):39–58.

Adner R, Puranam P, Zhu F (2019) What is different about digital strategy? From quantitative to qualitative change. *Strategy Sci.* 4(4):253–261.

Allcott H, Gentzkow M (2017) Social media and fake news in the 2016 election. *J. Econom. Perspect.* 31(2):211–236.

Bakos Y, Halaburda H (2020) Platform competition with multihoming on both sides: Subsidize or not. *Management Sci.* 66(12): 5599–5607.

Bamberger KA, Lobel O (2017) Platform market power. *Berkeley Technol. Law J.* 32(3):1051–1092.

Barnes J, Rosenberg M, Wong E (2020) China and Russia sow disinformation about how U.S. is handling the virus. *New York Times* (March 29).

Bechtold S, Tucker C (2014) Trademarks, triggers, and online search. *J. Empir. Leg. Stud.* 11(4):718–750.

Boudreau K (2012) Let a thousand flowers bloom? An early look at large numbers of software app developers and patterns of innovation. *Organ. Sci.* 23(5):1409–1427.

Boudreau KJ, Hagiu A (2009) Platform rules: Multi-sided platforms as regulators. Gawer A, ed. *Platforms, Markets and Innovation* (Edward Elgar Publishing, Cheltenham, UK), 163–191.

Bray J (2018) Automated fake account detection at LinkedIn. Accessed August 24, 2021, https://engineering.linkedin.com/blog/2018/09/automated-fake-account-detection-at-linkedin.

Breuer A, Eilat R, Weinsberg U (2020) Friend or faux: Graph-based early detection of fake accounts on social networks. *Proc. Web Conf. 2020* (Association for Computing Machinery, New York), 1287–1297.

Casadesus-Masanell R, Halaburda H (2014) When does a platform create value by limiting choice? *J. Econom. Management Strategy* 23(2):259–293.

Cennamo C, Santalo J (2013) Platform competition: Strategic trade-offs in platform markets. *Strategic Management J.* 34(11):1331–1350.

Cennamo C, Santalo J (2019) Generativity tension and value creation in platform ecosystems. *Organ. Sci.* 30(3):617–641.

Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: Synthetic minority over-sampling technique. *J. Artificial Intelligence Res.* 16:321–357.

Cho I-K, Kreps DM (1987) Signaling games and stable equilibria. *Quart. J. Econom.* 102(2):179–221.

Claussen J, Kretschmer T, Mayrhofer P (2013) The effects of rewarding user engagement: The case of Facebook apps. *Inform. Systems Res.* 24(1):186–200.

Connelly BL, Trevis Certo S, Duane Ireland R, Reutzel CR (2011) Signaling theory: A review and assessment. *J. Management* 37(1): 39–67.

Coy P (2021) Retail theft has gotten very organized. *New York Times* (December 1), https://www.nytimes.com/2021/12/01/opinion/retail-theft-ecommerce.html.

Cusumano MA, Gawer A, Yoffie DB (2021) Can self-regulation save digital platforms? *Indust. Corp. Change* 30:1259–1285.

Etter M, Ravasi D, Colleoni E (2019) Social media and the formation of organizational reputation. *Acad. Management Rev.* 44(1):28–52.

Fleder D, Hosanagar K (2014) Blockbuster culture's next rise or fall: The impact of recommender systems on sales diversity. *Management Sci.* 55(5):697–712.

Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. *J. Statist. Softw.* 33(1):1–22.

Fudenberg D, Tirole J (1991) *Game Theory* (MIT Press, Cambridge, MA).

Ghazawneh A, Henfridsson O (2013) Balancing platform control and external contribution in third-party development: The boundary resources model. *Inform. Systems J.* 23:173–192.

Gibbons R (1992) *Game Theory for Applied Economists* (Princeton University Press, Princeton, NJ).

Downloaded from informs.org by [193.60.238.99] on 07 April 2024, at 08:14 . For personal use only, all rights reserved.

Gleicher N, Ben Nimmo DA, Dvilyanski M (2021a) Adversarial threat report. Accessed December 13, 2021, https://about.fb.com/wp-content/uploads/2021/12/Metas-Adversarial-Threat-Report.pdf.

Gleicher N, Franklin M, Agranovich D, Ben Nimmo OB, Torrey M (2021b) Threat report the state of influence operations 2017–2020. Accessed December 8, 2021, https://about.fb.com/wp-content/uploads/2021/05/IO-Threat-Report-May-20-2021.pdf.

Goldfarb A, Tucker C (2019) Digital economics. *J. Econom. Lit.* 57(1):3–43.

Grinberg N, Joseph K, Friedland L, Swire-Thompson B, Lazer D (2019) Fake news on Twitter during the 2016 U.S. presidential election. *Science* 363(6425):374–378.

Gu G, Zhu F (2021) Trust and disintermediation: Evidence from an online freelance marketplace. *Management Sci.* 67(2):794–807.

Guess AM, Lerner M, Lyons B, Montgomery JM, Nyhan B, Reifler J, Sircar N (2020) A digital media literacy intervention increases discernment between mainstream and false news in the United States and India. *Proc. Natl. Acad. Sci. USA* 117(27):15536–15545.

Hamrick JT, Rouhi F, Mukherjee A, Feder A, Gandal N, Moore T, Vasek M (2018) The economics of cryptocurrency pump and dump schemes. Working paper, University of Tulsa, Tusla, OK.

Hand DJ (2009) Measuring classifier performance: A coherent alternative to the area under the ROC curve. *Machine Learning* 77:103–123.

Harris K (2015) The California Transparency in Supply Chains Act: A resource guide. Accessed December 8, 2021, https://oag.ca.gov/sites/all/files/agweb/pdfs/sb657/resource-guide.pdf.

Hastie T, Tibshirani R, Friedman J (2017) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer, New York).

Hsieh Y-Y, Vergne J-P (2023) The future of the web? The coordination and early-stage growth of decentralized platforms. *Strategic Management J.* 44:829–857.

Huang L, Joseph AD, Nelson B, Rubinstein BIP, Tygar JD (2011) Adversarial machine learning. *Proc 4th ACM Workshop Security Artificial Intelligence* (Association for Computing Machinery, New York), 43–58.

Huber TL, Kude T, Dibbern J (2017) Governance practices in platform ecosystems: Navigating tensions between cocreated value and governance costs. *Inform. Systems Res.* 28(3):563–584.

ICANN (2013) Beginner's guide to participating in ICANN. Accessed October 11, 2021, https://www.icann.org/en/system/files/files/participating-08nov13-en.pdf.

ICANN (2021) Intro to ICANN, ICANN Learn. Accessed October 11, 2021, https://learn.icann.org.

Jacobides MG, Lianos I (2021) Ecosystems and competition law in theory and practice. *Indust. Corporate Change* 30(5):1199–1229.

Karunakaran A, Orlikowski WJ, Scott SV (2022) Crowd-based accountability: Examining how social media commentary reconfigures organizational accountability. *Organ. Sci.* 33(1):170–193.

Kim A, Dennis AR (2019) Says who? The effects of presentation format and source rating on fake news in social media. *MIS Quart.* 43(3).

Kokkodis M, Lappas T, Kane GC (2022) Optional purchase verification in e-commerce platforms: More representative product ratings and higher quality reviews. *Production Oper. Management* 31(7):2943–2961.

Koo WW, Eesley CE (2020) Platform governance and the rural-urban divide: Sellers' responses to design change. *Strategic Management J.* 42:941–967.

Kovacic WE, Shapiro C (2000) Antitrust policy: A century of economic and legal thinking. *J. Econom. Perspect.* 14(1):43–60.

Kretschmer T, Leiponen A, Schilling M, Vasudeva G (2022) Platform ecosystems as meta-organizations: Implications for platform strategies. *Strategic Management J.* 43(3):405–424.

Kuan J, Lee G (2023) Governance strategy for digital platforms: Differentiation through information privacy. *Strategic Management Rev.* 4(2):161–191.

Mailath GJ, Okuno-Fujiwara M, Postlewaite A (1993) Belief-based refinements in signalling games. *J. Econom. Theory* 60(2):241–276.

Mas-Colell A, Whinston MD, Green JR (1995) *Microeconomic Theory*, vol. 1 (Oxford University Press, New York).

Mayzlin D, Dover Y, Chevalier J (2014) Promotional reviews: An empirical investigation of online review manipulation. *Amer. Econom. Rev.* 104(8):2421–2455.

Moravec PL, Kim A, Dennis AR (2020) Appealing to sense and sensibility: System 1 and system 2 interventions for fake news on social media. *Inform. Systems Res.* 31(3):987–1006.

Moravec PL, Minas RK, Dennis AR (2019) Fake news on social media: People believe what they want to believe when it makes no sense at all. *MIS Quart.* 43(4):1343–1360.

Murphy H, Reed J (2020) Facebook accuses telecoms groups of disinformation tactics: South-east Asian providers said to have used fake accounts to discredit rivals. *Financial Times* (February 12), https://www.ft.com/content/1096ad54-4d5f-11ea-95a0-43d18ec715f5.

Nimmo B, Torrey M (2022) Taking down coordinated inauthentic behavior from Russia and China. Accessed October 3, 2022, https://about.fb.com/wp-content/uploads/2022/09/CIB-Report_-China-Russia_Sept-2022-1-1.pdf.

Nix N, Ellison S (2023) Following Elon Musk's lead, big tech is surrendering to disinformation. *Washington Post* (August 25), https://www.washingtonpost.com/technology/2023/08/25/political-conspiracies-facebook-youtube-elon-musk/.

O'Mahony S, Ferraro F (2007) The emergence of governance in an open source community. *Acad. Management J.* 50(5):1079–1106.

Papanastasiou Y (2020) Fake news propagation and detection: A sequential model. *Management Sci.* 66(5):1826–1846.

Papernot N, McDaniel P, Jha S, Fredrikson M, Celik ZB, Swami A (2016) The limitations of deep learning in adversarial settings. *2016 IEEE Eur. Sympos. Security Privacy (EuroS&P)* (IEEE, Piscataway, NJ), 372–387.

Pennycook G, Rand DG (2019) Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proc. Natl. Acad. Sci. USA* 116(7):2521–2526.

Pennycook G, Bear A, Collins ET, Rand DG (2020) The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management Sci.* 66(11):4944–4957.

Persily N, Tucker JA (2021) How to fix social media? Start with independent research. *Brookings* (December 1), https://www.brookings.edu/research/how-to-fix-social-media-start-with-independent-research/.

Quiñonero-Candela J, Sugiyama M, Schwaighofer A, Lawrence ND, eds. (2008) *Data Set Shift in Machine Learning* (The MIT Press, Cambridge, MA).

Ricart JE, Snihur Y, Carrasco-Farré C, Berrone P (2020) Grassroots resistance to digital platforms and relational business model design to overcome it: A conceptual framework. *Strategy Sci.* 5(3):271–291.

Rietveld J, Schilling MA (2021) Platform competition: A systematic and interdisciplinary review of the literature. *J. Management* 47(6):1528–1563.

Rietveld J, Schilling MA, Bellavitis C (2019) Platform strategy: Managing ecosystem value through selective promotion of complements. *Organ. Sci.* 40(6):1232–1251.

Rietveld J, Seamans R, Meggiorin K (2021) Market orchestrators: The effects of certification on platforms and their complementors. *Strategy Sci.* 6(3):244–264.

Rosenberg M, Barnes J (2020) A bible burning, a Russian news agency and a story too good to check out. *New York Times*

Downloaded from informs.org by [193.60.238.99] on 07 April 2024, at 08:14 . For personal use only, all rights reserved.

(August 12), https://www.nytimes.com/2020/08/11/us/politics/russia-disinformation-election-meddling.html.

Ruokolainen T, Ruohomaa S, Kutvonen L (2011) Solving service ecosystem governance. *2011 15th IEEE Internat. Enterprise Distributed Object Computing Conf. Workshops* (IEEE, Piscataway, NJ), 18–25.

Schakowsky J (2021) Schakowsky introduces bill to protect consumers making online purchases. Accessed December 18, 2021, https://schakowsky.house.gov/media/press-releases/schakowsky-introduces-bill-protect-consumers-making-online-purchases.

Schmidt W, Buell R (2017) Experimental evidence of pooling outcomes under information asymmetry. *Management Sci.* 63(5):1586–1605.

S.H. 115-40 (2017) Disinformation: A primer in Russian measures and influence campaigns, Panel II. Accessed November 18, 2021, https://www.govinfo.gov/content/pkg/CHRG-115shrg25998/html/CHRG-115shrg25998.htm.

Shah SK (2006) Motivation, governance, and the viability of hybrid forms in open source software development. *Management Sci.* 52(7):1000–1014.

Shane S (2017) What intelligence agencies concluded about the Russian attack on the U.S. election. *New York Times* (January 6), https://www.nytimes.com/2017/01/06/us/politics/russian-hack-report.html.

Song M (2021) Estimating platform market power in two-sided markets with an application to magazine advertising. *Amer. Econ. J. Microeconom.* 13(2):35–67.

Tadelis S (2016) Reputation and feedback systems in online platform markets. *Annual Rev. Econom.* 8:321–340.

Timberg C (2016) Russian propaganda effort helped spread ‘fake news’ during election, experts say. *The Washington Post* (November 24), https://www.washingtonpost.com/business/economy/russian-propaganda-effort-helped-spread-fake-news-during-election-experts-say/2016/11/24/793903b6-8a40-4ca9-b712-716af66098fe_story.html.

Tiwana A, Konsynski B, Bush AA (2010) Research commentary—Platform evolution: Coevolution of platform architecture, governance, and environmental dynamics. *Inform. Systems Res.* 21(4): 685–697.

U.S. Court of Appeals for the Fifth Circuit (2023) State of Missouri; State of Louisiana; Aaron Kheriaty; Martin Kulldorff; Jim Hoft; Jayanta Bhattacharya; Jill Hines v. Joseph R. Biden, Jr.; Vivek H. Murthy; Xavier Becerra; Department of Health & Human Services; Anthony Fauci; et al. Case No. 23-30445, Document 238-1.

U.S. Department of Homeland Security (2019) *Combatting Targeted Disinformation Campaigns: A Whole-of-Society Issue* (Public-Private Analytic Exchange Program).

U.S. Office of the Director of National Intelligence (2017) Assessing Russian activities and intentions in recent US elections. U.S. Office of the Director of National Intelligence; ICA 2017-01D; Washington, DC (January 6).

Van Vlasselaer V, Eliassi-Rad T, Akoglu L, Snoeck M, Baesens B (2017) Gotcha! Network-based fraud detection for social security fraud. *Management Sci.* 63(9):3090–3110.

Vosoughi S, Roy D, Aral S (2018) The spread of true and false news online. *Science* 359:1146–1151.

Waldrop MM (2017) The genuine problem of fake news. *Proc. Natl. Acad. Sci. USA* 114(48):12631–12634.

Wang S, Pang M-S, Pavlou PA (2021) Cure or poison? Identity verification and the posting of fake news on social media. *J. Management* 38(4):1011–1038.

Wareham J, Fox PB, Josep LCG (2014) Technology ecosystem governance. *Organ. Sci.* 25(4):1195–1215.

Williamson E (2021) Alex Jones loses by default in remaining Sandy Hook defamation suits. *New York Times* (November 15), https://www.nytimes.com/2021/11/15/us/politics/alex-jones-sandy-hook.html.

Wilson T, Starbird K (2020) Cross-platform disinformation campaigns: Lessons learned and next steps. *Harvard Kennedy School Misinform. Rev.* 1(1):1–11.

Zhang Y, Li J, Tong TW (2019) Platform governance matters: How platform gatekeeping affects knowledge sharing among complementors. *Organ. Sci.* 30(6):1207–1231.

Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. *J. Roy. Statist. Soc. Ser. A* 67(2):301–320.

Anil R. Doshi is assistant professor of strategy and entrepreneurship at UCL School of Management. He received his doctorate from Harvard Business School. His research interests revolve around how organizations manage the digital information environment. His research contexts include involve generative AI, social media, and false information.

William Schmidt is an Associate Professor of Information Systems & Operations Management at Emory University. His research focuses on understanding and mitigating operational disruptions and applications of machine learning and game theory in operational decision making.