Expanding Platform Boundaries to Counter False Information

(Authors' names blinded for peer review)

False information, which can include disinformation, misinformation, and other 'fake news,' is persistent on digital content platforms. Research and strategies for constraining exposure to false information have focused on interrupting its spread within a platform, rather than its initial entry into a platform. We propose expanding the boundaries of the platform's functions to include monitoring and identifying false information internet domains. There are few burdens on establishing a domain, which has the unintended consequence of making it quick and inexpensive to create, replicate, and distribute false information through domains. To counter some of these advantages, we examine whether a platform can identify problematic domains *before* users engage with the domains' content on the platform. To establish the feasibility of such a strategy, we show that (1) a platform can predict false information domains at the time of registration, (2) the predictions are valuable under different platform governance types, and (3) the platform can exploit a false information creator's low transparency to sustain the efficacy of the prediction models.

1. Introduction

Disinformation, misinformation, and other 'fake news'—collectively false information—adversely impact the accurate representation of scientific knowledge (Scheufele and Krause 2019), democratic elections (H.R. 4617 2019), corporate activities (U.S. Department of Homeland Security 2019), and responses to global health emergencies (Gordon and Volz 2021). As a consequence, digital content platforms (hereafter, platforms) are under pressure to stem the proliferation of false information within their platforms (Kirkpatrick 2018). Doing so can be complicated, however, because false information can originate outside of the platform—most notably on internet domains—and then promoted on the platform by multiple automated or human users. Internet domain content is lightly regulated, and platforms receive little support interdicting such externally-originated false information. The Internet Corporation for Assigned Names and Numbers (ICANN), the organization responsible for policy and technical management of the internet's Domain Name System (DNS) and the closest thing to a governing body for the internet, holds that "internet governance should mimic the structure of the Internet itself—borderless and open to all" (ICANN 2013). In response to this challenge, we examine how platforms can evaluate domains and identify sources of false information before they penetrate the platform.

Access to traditional two-sided platforms can be limited via strict restrictions (Casadesus-Masanell and Halaburda 2014) or "soft" policies (Claussen et al. 2013). Similarly, content platforms can restrict access by identifying illegitimate users on their platforms at or near registration (Bray 2018, Breuer et al. 2020). For such intra-platform mitigation, the platform has control over the registration

process and the ability to actively screen suspect registrants by imposing additional registration steps or collecting non-registration data. There is little appetite to implement such controls for domain registrations, however.[1] This leaves domains as a common source of false information that can be used to infiltrate other platforms. While platforms cannot shut down the source of externally generated false information, they can target content after it has entered the platform (Candogan and Drakopoulos 2020, Papanastasiou 2020), or reduce its consumption by tagging and fact-checking (Moravec et al. 2020) or highlighting sources (Kim and Dennis 2019). But identifying and intervening against content only after it appears on the platform can entail delay and provide opportunities for content to propagate within the platform or avoid detection altogether.

Defining and adjusting the scope of a firm's boundaries, activities, or interests is a canonical topic in strategy research (Williamson 1999), and expanding content monitoring activities beyond the traditional confines of the platform aligns with an emerging view in the strategy literature of platforms as meta-organizations (Kretschmer et al. 2020). We argue that platforms can proactively assess content on domains outside the platform's boundaries, and use its assessment to support intra-platform interdiction efforts, before links or content from those domains penetrates the platform. To achieve this, we employ machine learning algorithms as a means to generate economical predictions (Agrawal et al. 2019). We establish that a platform can provide predictions of external content that are (1) effective, (2) flexible to different governance priorities, and (3) robust to changes in the domain registrant's behavior. To assess efficacy, we assemble a set of general domains and known false information domains, collect each domain's original registration data, and use the combined data to train an elastic net classifier using k-fold cross validation. Even with limited data and a standard machine learning algorithm, the results are encouraging—the area under the curve (AUC) of the classifier's receiver operating characteristic (ROC) is 0.922. We then model the value of the predicted outcomes to show how the predicted probabilities can be employed by platforms with different governance priorities. Finally, we use a signaling game to assess the strategic actions of the registrants, and show that three equilibrium outcomes are sustained—a non-distortive separating equilibrium, a distortive separating equilibrium, and a pooling equilibrium. All three outcomes disadvantage false information registrants and weakly advantage general registrants.

2. Machine Learning Classifier

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 sample of 20% of the data for the test stage to

[1] ICANN states that it "does not control content on the Internet. It cannot stop spam and it does not deal with access to the Internet." (ICANN Learn 2021)

estimate the performance of the classifiers. The other 80% of the data is used to train and validate the classifier. Second, we use an elastic net penalized logit regression algorithm to develop our classifiers (Zou and Hastie 2005). The algorithm is similar to the maximum likelihood estimator, but it employs two hyperparameters, α and λ, to impose a compound penalty from ridge and Lasso regressions on the feature weights in the final classifier. The first hyperparameter, $0 \leq \alpha \leq 1$, governs the mix between the two penalties. The second hyperparameter, $\lambda > 0$, controls the extent to which coefficients are penalized in the algorithm. We set $\alpha = 0.99$, which is essentially a Lasso classifier without erratic behaviors that arise from highly correlated variables (Friedman et al. 2010). We select λ using k-fold cross-validation (kFCV). In the Online Appendix, we provide details for the classifier development process.

2.1. Sample and Feature Extraction

Our sample of false information domains comes from Allcott and Gentzkow (2017). They collected links to articles that were proven to be false by Snopes, Politifact, or Buzzfeed, in the time surrounding the November 2016 U.S. presidential election. The database consists of articles 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.[2] We drop eight more domains that were sub-domains on aggregators or content creating platforms, namely Pocket, Youtube, Wordpress, and Blogspot. We augment these observations with a sample of 4,000 general purpose domains provided by DomainTools, an online security and data company. These domains were registered over the same time period as the false information domains (see the Online Appendix for details). We drop ten domains that were missing registration date information or were registered outside our period of interest. Our final dataset consists of 3,990 general and 354 false information domains that we identify using a binary measure for false information ("1") or not ("0"). To address the imbalance between the classes, we use Synthetic Minority Over-Sampling Technique, which 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 obtain each domain's registration information from DomainTools. 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. From this, we extract 1,139 features to train the algorithm. In the Online Appendix, we provide details of the feature engineering process.

[2] 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.

2.2. General Results

The resulting elastic net classifier has non-zero weights on 193 features pertaining to the domain name's characteristics, geography and privacy of the registrant, and characteristics of the domain extension and registrar. The classifier generates a predicted probability for each domain that the domain will produce false information in the future. By applying a cutoff threshold to the predicted probabilities, we can generate a predicted class for each domain. Figure 1 presents the receiver operating characteristic (ROC) curve, which shows the tradeoff between the sensitivity (or true positive rate) and one minus specificity (or false positive rate) at thresholds varying from 0 to 1. The ROC's area under the curve (AUC) of 0.922 can be compared to 0.500, which is the AUC of a random classifier (denoted by the dashed line). In order to assess whether the strong predictive power arises from the data or classifier design choices, we run a series of alternative algorithms (see the Online Appendix). The results are consistent with the performance of our presented classifier.

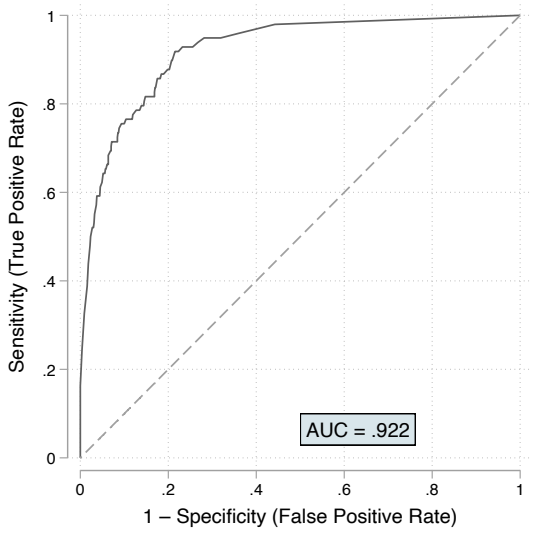


Figure 1 **Prediction performance for elastic net classifier of false information domains. Hyphenated line represents performance of a random classifier.**

2.3. Optimizing Results Using Platform Priorities

A platform can choose its optimal threshold by solving an optimization problem that balances the economic costs and benefits associated with the four outcomes: true positives (TP), false positives (FP), true negatives (TN), false negatives (FN). This economic modeling exercise can be tailored to the governance policies of the platform. To illustrate this, we employ a simple economic model of the prediction's value for three platform governance types: *strict, moderate, and lax*. Our economic

model extends the H-measure approach of assessing classifier performance (Hand 2009) by including the economic implications of true classifications, when appropriate.

The platform incurs a benefit V_s, for identifying a false information domain. The platform's policy is to put each predictive positive domain through additional verification at a cost C_v. Each false positive imposes a cost C_p, arising from frictions caused by incorrectly identifying an otherwise legitimate domain, and a cost C_s, for misidentifying a false information domain.[3] The platform captures these costs in a model, such as Equation 1, that aligns with its governance policies. The model serves as an objective function in an optimization problem to determine the platform's optimal threshold using the test data. The platform then classifies newly or recently registered domains by applying that threshold to the probabilities generated by the machine learning model for new domains.

$$\text{Value} = V_s(TP) - [(C_v(TP+FP) + C_p(FP) + C_s(FN)] \tag{1}$$

Table 1 summarizes our results for the three governance types (see the Online Appendix for details). We normalize $V_s = 1$ and compute the value function for a range of values of C_v, C_p, and C_s. For each type, we compute the value of the machine learning model using an optimal threshold from Equation (1) and compare it to that type's highest value from among three alternatives prediction strategies: assume no false information domains, assume all false information domains, or randomly classify domains. The strict platform has a low C_v and C_p, and a high C_s, relative to the value of correctly identifying a false information provider. The optimal thresholds and the incremental values (presented as unitless values per 1,000 domains) of the machine learning classifier differ markedly among the platform types. Intuitively, the optimal threshold increases from strict (0.10) to moderate (0.54) to lax (0.86). However, the incremental value of the machine learning classifier based on those optimal thresholds does not change monotonically from strict (117.6 per 1,000 domains) to moderate (119.8) to lax (47.3). As opposed to tuning machine learning classifiers to an accuracy measure, these observations underscore the importance of tuning machine learning classifiers to a value measure that reflects the priorities of the organization that the classifier serves.

3. Analytical Signaling Model

Machine learning classifiers are largely theory-free and offer little framework to interpret the results. To address this, and to lay a foundation for future empirical study, we develop and analyze a minimum viable dynamic game of incomplete information between a risk neutral monitoring function within a platform (denoted M, he/him) and a domain registrant (denoted R, she/her). The monitor evaluates domains at the time of registration, but he does not have control over them and cannot

[3] True negatives do not influence the economic model in this example, but this assumption is easily relaxed.

Table 1 **Confusion matrices for three representative platform governance types.**

	Strict	Moderate	Lax
V_s	1.0	1.0	1.0
C_v	0.1	0.2	0.4
C_p	0.1	0.2	0.4
C_s	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
Recall / sensitivity	0.918	0.755	0.592
False positive rate	0.215	0.093	0.038
Precision	0.352	0.507	0.667
Accuracy	0.800	0.890	0.921
F_1	0.508	0.607	0.627
Incremental value (per 1,000 domains)	117.6	119.8	47.3

True positives, true negatives, false positives, and false negatives are a proportion of all predictions.

prevent them from operating. The registrant has private information on the type of domains that she wishes to establish, and can be one of two types, $t \in \{L, H\}$. Type H has high legitimacy, and is interested in establishing a domain that is free of deception. Type L has low legitimacy, and is interested in establishing a domain that contains deception, such as producing false information. It is common knowledge that the registrant is type H with probability $h \in (0, 1)$ and type L with probability $(1 - h)$. The registrant has some discretion on the composition of her registration information, i.e. how she completes the registration and the level of detail she provides. Subtle composition differences between the registrant types may provide information about their type.

For expositional clarity, we use the registrant's transparency T to describe the signaling mechanism, although any feature can be used which poses a differential cost to L-type and H-type registrants. Scholars have employed transparency as a mechanism in a variety of settings, including its influence on the behavior of competing platforms (Li and Zhu 2021), firm investments in corporate social responsibility initiatives (Wu et al. 2020), and service operations decisions on platforms (Mejia and Parker 2021). In our context, transparency measures the degree to which registration data provides verifiable information to identify the registrant's true identify. More (less) transparency makes it easier (harder) to attribute a domain to a person or entity. An H-type registrant receives some operational value from transparency (for instance, by facilitating information flows with the registrar), although this value need not be monotonically increasing (for instance, due to privacy concerns). An L-type registrant, however, views transparency as undesirable since it can reveal her true identity and expose her to operational costs (sanction or legal action). As a result,

transparency has a dual effect on the registrant's utility—an operational impact and a signal of her type that can yield differential monitoring by the platform. To facilitate a clear understanding of these effects, we include them separately in the model.

3.1. Utility Functions

Controlling a domain provides value $V(t)$ to the t-type registrant. The operational benefit (or cost, if negative) of transparency is $O(t,T)$. In line with standard signaling game models, we assume that both the cost and marginal cost of transparency are lower for the H-type registrant compared to the L-type registrant (Mas-Colell et al. 1995). Since O represents a benefit, $O(H,T) > O(L,T)$ and $\frac{\partial O(H,T)}{\partial T} > \frac{\partial O(L,T)}{\partial T}$ $\forall T$. Upon assessing the registrant's signal, the monitor may update his belief that the registrant is type H with probability $\theta \in [0,1]$ and type L with probability $(1-\theta)$. Based on his updated beliefs, the monitor assigns a monitoring level to the registrant. The cost $M(t,\theta)$ that this monitoring imposes on the registrant decreases in θ and can even be negative (a benefit) at high levels of θ. Monitoring cost can take many forms in practice, including assigning a certification, subjecting the registrant's domain to increased oversight, or publishing a warning about the registrant's domain. Similar actions can be seen in practice when social media companies label false or misleading content, or internet security companies publish lists of phishing domains.

The registrant's utility function is expressed as:

$$U_R(t,T,\theta) = V(t) - M(t,\theta) + O(t,T). \tag{2}$$

As a function within the platform, the monitor's decision is to set a monitoring level $A \in [0,1]$ for the registrant, and his utility function depends on the error of this assessment. We capture the form of his utility function following a simple structure for risk neutral actors in Gibbons (1992).

$$U_M(T,\theta) = -\left(A - (1-\theta)\right)^2. \tag{3}$$

This reflects his desire to assign a monitoring level to the registrant that is commensurate with his perceived likelihood that the registrant will publish false information, so $A = 1-\theta$. In equilibrium, the monitor's updated belief resolves to three cases: the registrant is an H-type $(\theta = 1)$, the registrant is an L-type $(\theta = 0)$, or the monitor maintains his prior belief $(\theta = h)$.

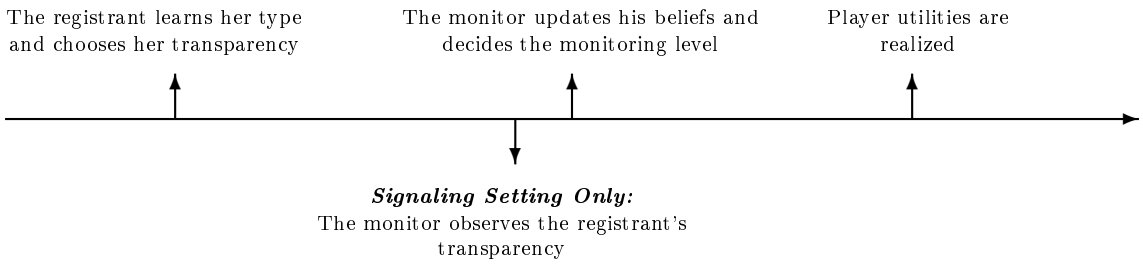
3.2. Analysis

We use our analytical model to examine how behavior and utilities differ between a non-signaling setting and a signaling setting. Supporting technical details are provided in the Online Appendix. In the non-signaling setting, none of the players realize that the registrant's level of transparency can provide information about the registrant's type, so the monitor does not change his beliefs after observing the registration information. This setting provides insight into player behavior leading

to the performance of our machine learning classifiers. To determine whether this performance can continue once all of the players recognize the signaling power of registration information, we apply our model to the signaling setting in which both players know that registration information can provide a signal of the registrant's type.

Figure 2 summarizes the sequence of events. First, the registrant learns her type and chooses the amount of transparency that she will provide through her platform registration information. The monitor observes the registrant's transparency (if applicable), updates his beliefs, and applies a monitoring level to the registrant. Finally, both players realize their utilities.

Figure 2 Timeline of the Model



To keep our model simple, we do not explicitly model a societal benefit. Instead, one can assume that societal benefits increase with the H-type's utility and decrease with the L-type's utility. To more readily represent the main insights of our model structure, we generate analytical results assuming that $O(H,T)$ is first increasing and then decreasing in T; $O(L,T)$ is decreasing in T; and $T \in \mathbb{R}^+$. Relaxing these assumptions can yield additional insights, as we briefly illustrate in Section 3.2.3.

3.2.1. Non-Signaling Setting. In the non-signaling setting, the monitor does not update his beliefs based on the registrant's level of transparency, so transparency has no impact on the registrant's monitoring level, and neither registrant type has an incentive to mimic the other type. We formally describe the registrant's equilibrium transparency level in Lemma 1.

Lemma 1 *If registration information is not recognized as a signal of the registrant's type, an L-type registrant's transparency is*

$$T_L^* = \underset{T}{\operatorname{argmax}}\, O(L,T), \tag{4}$$

and an H-type registrant's transparency is

$$T_H^* = \underset{T}{\operatorname{argmax}}\, O(H,T). \tag{5}$$

The monitor's updated beliefs are $\theta = h$.

The resulting utility for the L-type is $U_R(L,T,\theta) = V(L) - M(L,h) + O(L,T_L^*)$ and for the H-type is $U_R(H,T,\theta) = V(H) - M(H,h) + O(H,T_H^*)$.

3.2.2. Signaling Setting. We next analyze the equilibrium outcomes when both player's recognize that registration information transparency provides a signal of the registrant's type.[4] 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) prioritize the H-type registrant's preferred signal, and conditional on that, prioritize 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 rather than the opposite. In addition to providing reasonable and intuitive results, an LMSE yields a unique prediction in our setting.

The L-type has an incentive to reduce her monitoring costs by mimicking the H-type. This can induce the H-type to respond by overinvesting in the signal in order to distinguish herself as an H-type. Although the L-type has an incentive to mimic the H-type, she will not choose a level of transparency that is dominated by T_L^*, i.e.:

$$U_R(L,T,\theta=1) \leq U_R(L,T_L^*,\theta=0). \tag{6}$$

We identify $T' = \text{argmin}_T[U_R(L,T,\theta=1) - U_R(L,T_L^*,\theta=0)]^2 \ \forall T > T_L^*$, i.e. the lower bound of all transparency levels greater than T_L^* that satisfy inequality (6). To separate from the L-type, the H-type can choose a transparency level, $T^s = max\{T', T_H^*\}$, that is strictly dominated for the L-type and, conditional on that, best for her own utility. If $T' < T_H^*$, the H-type can choose her optimal level of transparency, T_H^*, and still credibly reveal her type to the monitor. If $T' > T_H^*$, the H-type must deviate from T_H^* in order to credibly reveal her type, but she will not choose a level of transparency that is dominated by T_H^* under a weighted belief, i.e.:

$$U_R(H,T,\theta=1) \leq U_R(H,T_H^*,\theta=h). \tag{7}$$

We identify $T'' = \text{argmin}_T[U_R(H,T,\theta=1) - U_R(H,T_H^*,\theta=h)]^2 \ \forall T > T_H^*$, i.e. the lower bound of all transparency levels greater than T_H^* that satisfy inequality (7). If $T' \geq T''$, the H-type has no

[4] It can be shown that the results are the same if only the registrant recognizes the signal. If only the platform recognizes the signal, he will assign a differential level of monitoring according to each player's true type, thereby decreasing the monitoring cost of the H-type while increasing the monitoring cost of the L-type. The registrant may learn the signal mechanism in a repeated game, which will lead to our studied equilibrium.

incentive to differentiate herself from the L-type, and she will instead choose a pooling transparency level that is best for her, given that an L-type will mimic it.

We summarize each player's strategy in Lemma 2.

Lemma 2 *In the signaling setting, each player's equilibrium strategy is:*

(i) The L-type registrant provides transparency:

$$T = \begin{cases} T_L^* & \text{if } T' < T_H^*, \\ T_L^* & \text{if } T_H^* \leq T' < T'', \\ T_H^* & \text{if } T' \geq T''. \end{cases} \tag{8}$$

(ii) The H-type registrant provides transparency:

$$T = \begin{cases} T_H^* & \text{if } T' < T_H^*, \\ T' & \text{if } T_H^* \leq T' < T'', \\ T_H^* & \text{if } T' \geq T''. \end{cases} \tag{9}$$

(iii) The monitor assigns a monitoring level $1-\theta$, and her posterior beliefs, θ, are:

$$\theta = \begin{cases} 0 & \text{if } T < T_H^*, \\ h & \text{if } T_H^* \leq T < T^s, \\ 1 & \text{if } T \geq T^s \geq T_H^*. \end{cases} \tag{10}$$

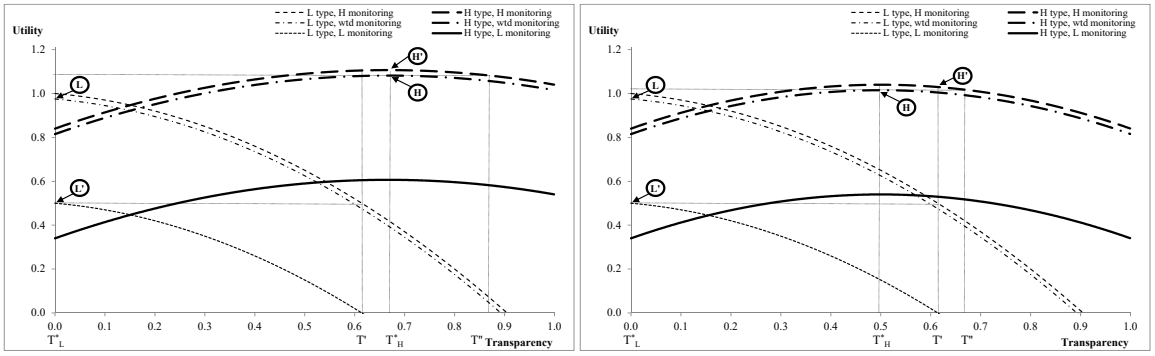
There are three equilibrium outcomes in our analysis. In a non-distortive separating equilibrium, the registrant chooses her preferred transparency from Lemma 1 based on her type (an L-type's choice is T_L^* and an H-type's choice is T_H^*), and the monitor correctly infers the registrant's type. In a distortive separating equilibrium, the L-type registrant chooses T_L^*, the H-type chooses more than her preferred transparency with T', and the monitor correctly infers the registrant's type. In a pooling equilibrium, the L-type registrant chooses more than her preferred transparency T_H^*, the H-type chooses her preferred transparency T_H^*, and the monitor maintains her prior beliefs of the registrant's type. Note that in equilibrium, θ is weakly increasing in T. Only the L-type registrant provides transparency that is lower than T_H^* and only the H-type registrant provides transparency that is higher than T_H^*.

3.2.3. Outcomes. We formalize the difference in the utility outcomes across the non-signaling and signaling settings for each type in the following proposition.

Proposition 1 *In the signaling setting, the L-type registrant's utility is strictly lower and the H-type registrant's utility is weakly higher than the respective utilities under a non-signaling setting.*

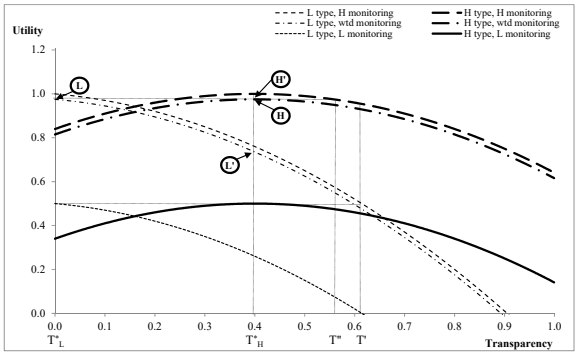
The non-signaling game provides the best outcome for the L-type registrant—low operational impact and modest monitoring costs. In the signaling game, however, either the operational impact or the monitoring costs will increase for the L-type. In contrast, the H-type's outcome in the non-signaling game can be improved in the signaling game through lower monitoring costs. Figure 3

illustrates these effects with representative comparisons of the transparency choices and utilities for the two types across the two settings. Each subfigure captures a representative equilibrium: non-distortive separating equilibrium (Figure 3(a)), distortive separating equilibrium (Figure 3(b)), and pooling equilibrium (Figure 3(c)). In all three figures, L (H) points to the L-type's (H-type's) transparency and utility outcomes in the non-signaling setting, and L' (H') points to the L-type's (H-type's) transparency and utility outcomes in the signaling setting.



(a) Non-distortive separating

(b) Distortive separating



(c) Pooling

Figure 3 **Representative Equilibrium Outcomes. H and L identify the transparency choices and utilities in the non-signaling setting for the H-type and L-type registrants, respectively. H′ and L′ identify the transparency choices and utilities in the signaling setting under a non-distortive separating equilibrium (Panel a), a distortive separating equilibrium (Panel b), and a pooling equilibrium (Panel c). We use $V(L) = 1$, $V(L) = 0.84$, $M(L,\theta) = M(H,\theta) = (1-\theta)0.5$, and $O(L,T) = -0.2T - T^2$. The only difference is that $O(H,T) = 0.8T - 0.6T^2$ in Panel a, $O(H,T) = 0.8T - 0.8T^2$ in Panel b, and $O(H,T) = 0.8T - T^2$ in Panel c.**

Relaxing our modeling assumptions can further inform how monitors can constrain illegitimate

actors and expand the generalizability of the model. For instance, we assumed that $T \in \mathbb{R}^+$. In practice however, registration processes often implicitly limit the amount of transparency that can be provided. Such limits may constrain the ability of legitimate registrants from voluntarily delineating themselves from illegitimate registrants, leading to a pooling outcome.[5] Accounting for transparency constraints can inform how a platforms can deliver more value to legitimate domains in exchange for participating in robust authentication of registration information.[6] An example of this in the non-digital world is the PreCheck® service offered by the Transportation Security Administration (TSA). In this case, people willingly pay a processing fee and provide verifiable information to the TSA in exchange for expedited clearance when boarding flights in the United States. Similarly, the platform could offer a certification to positively influence a domain registrant's participation in more active verification processes (Rietveld et al. 2021). Targeted interventions can have the additional benefit of increasing the overall value of the platform's ecosystem (Rietveld et al. 2019).

4. Implications and Conclusions

We propose that digital content platforms should look outside the confines of the platform to assess external sources of illegitimate content. In the physical world, this is loosely akin to the common business practice of performing quality checks on goods at a supplier's facility (referred to as "pre-shipment inspection") rather than dealing with defective goods after they are delivered. The parallels and differences with this simple analogy open up several avenues for future research. For instance, just as quality control of suppliers can be outsourced to an independent firm, evaluating domains could be performed by an independent monitor. An independent monitor could serve multiple platforms, add transparency, avoid redundant monitoring costs, and streamline interactions with those being monitored (standard setting, resolving false predictions, etc.). By showing in Section 2.3 that the monitoring function is valuable under different governance regimes, we offer some indication that an independent monitor could serve multiple platforms if it provides predicted probabilities rather than classes. An external monitor may still introduce knowledge and prioritization frictions, however, especially since the domain predictions will integrate with the platform's intra-platform interdiction efforts. We leave a deeper examination of this organizational structure, potentially through the lens of maintaining and transferring knowledge within and across firms (Kogut and Zander 1992, Grant 1996), to future research.

While we establish that monitoring external information sources is feasible and can add value to a platform's monitoring effort (Helper et al. 2021), there are opportunities to expand on the

[5] To see this in the model, consider when the maximum value of T is below T' and $T' < T''$.

[6] Allowing registrants to provide greater transparency can yield a non-distortive or distortive separating equilibrium, which can provide a higher utility for and H-type compared to a pooling equilibrium.

monitoring function. Our proof-of-concept machine learning classifier uses limited data, a relatively simple algorithm, and a small sample size compared to the daily volume of registered domains.[7] Any of these dimensions could open interesting lines of research, including the scope, scale, and constraints of external monitoring. For instance, even if the proportion of false positives and false negatives is small, their number can be large in practical settings. This highlights the need for a staged validation and escalation process that combines artificial intelligence and human intelligence (Teodorescu et al. 2021). Integrating these roles and organizing them across firm boundaries is salient in many organizational contexts as the use of artificial intelligence continues to expand.

Finally, any effort to disrupt the flow of false information must be adaptable and expand beyond the roles outlined in this paper. While we focus on the initial domain registration, platforms should also monitor domain transfers to account for registrants who purchase an existing domain. As in other settings, illegitimate actors in our setting will seek to defeat countermeasures in unanticipated ways. Effectively overcoming such challenges calls for creative, boundary spanning solutions and research.

References

Agrawal, Ajay, Joshua Gans, Avi Goldfarb. 2019. Prediction, judgment, and complexity: A theory of decision-making and artificial intelligence. Ajay Agrawal, Joshua Gans, Avi Goldfarb, eds., *The Economics of Artificial Intelligence: An Agenda*. University of Chicago Press, 89–114.

Allcott, Hunt, Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of Economic Perspectives* **31**(2) 211–236.

Bray, Jenelle. 2018. Automated fake account detection at LinkedIn. URL `https://engineering.linkedin.com/blog/2018/09/automated-fake-account-detection-at-linkedin`. Accessed Aug 24, 2021.

Breuer, Adam, Roee Eilat, Udi Weinsberg. 2020. Friend or faux: Graph-based early detection of fake accounts on social networks. *Proceedings of The Web Conference 2020*. 1287–1297.

Candogan, Ozan, Kimon Drakopoulos. 2020. Optimal signaling of content accuracy: Engagement vs. misinformation. *Operations Research* **68**(2) 497–515.

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

Chawla, Nitesh V, Kevin W Bowyer, Lawrence O Hall, W Philip Kegelmeyer. 2002. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* **16** 321–357.

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

[7] Around 200,000 domains are registered every day, `https://dnpedia.com/tlds/daily.php`.

Friedman, Jerome, Trevor Hastie, Rob Tibshirani. 2010. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33**(1) 1.

Fudenberg, Drew, Jean Tirole. 1991. *Game Theory*. MIT Press, Cambridge, Mass.

Gibbons, Robert. 1992. *Game theory for applied economists*. Princeton University Press.

Gordon, Michael, Dustin Volz. 2021. Russian disinformation campaign aims to undermine confidence in Pfizer, other Covid-19 vaccines, U.S. officials say. *The Wall Street Journal* (March 7).

Grant, Robert M. 1996. Toward a knowledge-based theory of the firm. *Strategic Management Journal* **17**(S2) 109–122.

Hand, David J. 2009. Measuring classifier performance: A coherent alternative to the area under the roc curve. *Machine Learning* **77** 103–123.

Helper, Susan, Raphael Martins, Robert Seamans. 2021. Who profits from Industry 4.0? Theory and evidence from the automotive industry. *Working paper* .

H.R. 4617. 2019. Stopping Harmful Interference in Elections for a Lasting Democracy Act. 116th Congress.

ICANN. 2013. Beginner's guide to paticipating in ICANN. https://www.icann.org/en/system/files/files/participating-08nov13-en.pdf.

ICANN Learn. 2021. Intro to ICANN, ICANN Learn. https://learn.icann.org.

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

Kirkpatrick, David D. 2018. 'Fake news' investigators rebuke Facebook. *The New York Times* July 29.

Kogut, Bruce, Udo Zander. 1992. Knowledge of the firm, combinative capabilities, and the replication of technology. *Organization Science* **3**(3) 383–397.

Kretschmer, Tobias, Aija Leiponen, Melissa Schilling, Gurneeta Vasudeva. 2020. Platform ecosystems as meta-organizations: Implications for platform strategies. *Strategic Management Journal, articles in advance* .

Li, Hui, Feng Zhu. 2021. Information transparency, multihoming, and platform competition: A natural experiment in the daily deals market. *Management Science* **67**(7) 4384–4407.

Mailath, George J., Masahiro Okuno-Fujiwara, Andrew Postlewaite. 1993. Belief-based refinements in signalling games. *Journal of Economic Theory* **60**(2) 241–276.

Mas-Colell, Andreu, Michael Dennis Whinston, Jerry R Green, et al. 1995. *Microeconomic Theory*, vol. 1. Oxford University Press New York.

Mejia, Jorge, Chris Parker. 2021. When transparency fails: Bias and financial incentives in ridesharing platforms. *Management Science* **67**(1) 166–184.

Moravec, Patricia L, Antino Kim, Alan R Dennis. 2020. Appealing to sense and sensibility: System 1 and system 2 interventions for fake news on social media. *Information Systems Research* **31**(3) 987–1006.

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

Rietveld, Joost, Melissa A. Schilling, C. Bellavitis. 2019. Platform strategy: Managing ecosystem value through selective promotion of complements. *Organization Science* **40**(6) 1232–1251.

Rietveld, Joost, Robert Seamans, Katia Meggiorin. 2021. Market orchestrators: The effects of certification on platforms and their complementors. *Strategy Science* **6**(3) 244–264.

Scheufele, Dietram A., Nicole M. Krause. 2019. Science audiences, misinformation, and fake news. *Proc Natl Acad Sci* **116**(16) 7662–7669.

Teodorescu, Mike H. M., Lily Morse, Yazeed Awwad. 2021. Failures of fairness in automation require a deeper understanding of human-ml augmentation. *MIS Quarterly* **45**(3) 1483–1499.

U.S. Department of Homeland Security. 2019. Combatting targeted disinformation campaigns: A whole-of-society issue. *Public-Private Analytic Exchange Program.*

Van Vlasselaer, Véronique, Tina Eliassi-Rad, Leman Akoglu, Monique Snoeck, Bart Baesens. 2017. Gotcha! network-based fraud detection for social security fraud. *Management Science* **63**(9) 3090–3110.

Williamson, Oliver E. 1999. Strategy research: Governance and competence perspectives. *Strategic Management Journal* **20**(12) 1087–1108.

Wu, Yue, Kaifu Zhang, Jinhong Xie. 2020. Bad greenwashing, good greenwashing: Corporate social responsibility and information transparency. *Management Science* **66**(7) 3095–3112.

Zou, Hui, Trevor Hastie. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society* **67**(2) 301–320.