Editorial

Machine Learning for Consumer Forensics (or We Need to Talk About Online Enforcement of Regulations)

In 2025, the new European Commission will take office dealing inter-alia with topics on tech sovereignty, security, and democracy. Finland's Henna Virkkunen has been appointed Executive Vice President to lead these pivotal areas,[1] overseeing Europe's digitalisation strategy through the Digital Decade and fostering innovation in key sectors such as AI, cloud computing, and quantum technologies. Her mandate also includes ensuring swift and effective enforcement of the Digital Services Act (DSA) and the Digital Markets Act (DMA) as required. Additionally, there is a clear commitment to addressing unethical online practices, with particular emphasis on dark patterns, influencer marketing, addictive design, online profiling, hamful content (including misinformation and cyberbullying). As an interdisciplinary data scientist, I view these developments with great anticipation for the advancements ahead.

Viewing consumer vulnerabilities as a technical problem (addressing monitoring and enforcement) requires three main steps: measurement and data collection, building detection models and finally evaluating the approach. All these together operationalize what Goanta has described as "consumer forensics".[2] This editorial outlines how the typical data science/machine learning pipeline can help address these challenges.

I. Data Collection: Measuring prevalence

Data collection is the first and most crucial step in building an effective consumer forensics framework. Monitoring the digital landscape for vilations related to regulations like DSA —whether it is dark patterns/addictive design, influencer marketing, or harmful content—requires vast datasets that will help measure the impact of the studied problem, act as evidence/proof of potential violations and provide the foundation for building any model (either for understanding patterns or predicting violations). These datasets used to be generated by scraping websites, tracking user behavior patterns, monitoring social media interactions. One of the most anticipated actions in 2025 would be the implementation of DSA in relation to data access for researchers, as well as, the evolution of EC databases such as the DSA Transparency Database or the Digital Services Terms

---

1 European Commission, *Mission letter to commissioner Henna Virkunnen* <https://commission.europa.eu/document/3b537594-9264-4249-a912-5b102b7b49a3_en> accessed on 26 September 2024.
2 C Goanta, 'Digital Detectives: A Research Agenda for Consumer Forensics' (2023) 8(3) European Papers-A Journal on Law and Integration 647.

and Conditions Database. The DSA transparency database has been analyzed[3] as a first effort to uncover moderation efforts of platforms, but systematic access to comprehensive data is still challenging for researchers.[4]

Tracking influencer marketing requires a diverse set of data. Influencers frequently promote products through sponsored posts, sometimes without proper disclosures, violating advertising regulations. Machine learning models can analyze content from social media posts (text, images, and videos) to detect undisclosed paid promotions.[5] Natural Language Processing (NLP) frameworks can then be employed to analyze captions for keywords such as "ad," "sponsored," or other disclosure terms, while image recognition can flag product placements. Longitudinal collection of such data[6] allows the observation of evolution of influencer marketing as well as detecting inflection points on social media platform affordances. At the same time, collecting data across platforms and languages[7] is essential for understanding the ecosystem of digital advertising and collecting the necessary evidence.

Protecting consumers from harmful content online has been the "hot potato" of platform content moderation. The need for platforms to moderate content according to the law and not to their arbitrary terms and services is the promise that comes along with the DSA. Meanwhile, it is essential to perform research on such sensitive topics (which it is highly anticipated to be improved by DSA standards). For example, data collection for cyberbullying involves scanning forums, social media comments and online communities for abusive language or content that violates the relevant laws. At the same time, concepts like "cancel culture" (holding individuals accountable for actions that are perceived as harmful) can result in public shaming, boycotts, exclusion and eventually escalate to personal attacks, harassment, effectively crossing the line into cyberbullying. In previous research[8] we have constucted a framework on how to measure, identify and evaluate cancel culture, while delivering the first dataset of that kind.

Dark patterns and addictive design lie in the heart of how modern websites and platforms operate. Certain design elements (e. g. endless scrolling, reward mechanisms) intentionally keep consumers engaged beyond what is reasonable. Recent research efforts have constructed and delivered datasets for dark patterns[9] both in web and mobile environments. The increasingly adaptive design of websites and mobile apps, the complexity of user interfaces and the high degree of personalization adds extra obstacles into a standardized data collection of dark patterns.

Moreover, the process of gathering such vast datasets poses several challenges, particularly in terms of access, privacy, and cross-platform consistency. The Digital Services Act (DSA) can play a pivotal role in addressing these issues by ensuring standardized access to platform data for researchers, while also enforcing transparency in content moderation and advertising practices. By mandating clear data-sharing protocols, the DSA will help bridge the current gaps in data availability, thus supporting a more comprehensive understanding of the digital landscape and its risks for consumers.

3   R Kaushal, J Van De Kerkhof, C Goanta, G Spanakis and A Iamnitchi, 'Automated Transparency: A Legal and Empirical Analysis of the Digital Services Act Transparency Database' in *Proceedings of The 2024 ACM Conference on Fairness, Accountability, and Transparency* (ACM 2024).
4   D Halil, K Kollnig and A Tamò-Larrieux, 'Regulating pressing systemic risks – but not too soon? Comparative Analysis of the Implementation of Data Access Requests to Platform Data under Article 40(4) of the EU Digital Services Act <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4959049 > accessed on26 September 2024.
5   S Kim, JY Jiang, M Nakada, J Han, J. and W Wang, 'Multimodal post attentive profiling for influencer marketing' in Y Huang, I King, *WWW '20: Proceedings of The Web Conference 2020* (ACM 2020).
6   T Bertaglia and C Goanta and G Spanakis and A Iamnitchi, 'Influencer Self-Disclosure Practices on Instagram: A Multi-Country Longitudinal Study' < https://arxiv.org/abs/2407.09202 > accessed 26 September 2024.
7   H Gui, T Bertaglia, C Goanta, S de Vries and G Spanakis 'Across Platforms and Languages: Dutch Influencers and Legal Disclosures on Instagram, YouTube and TikTok' In *Proceedings of the International Conference on Advances in Social Network Analysis and Mining (ASONAM 2024)* (forthcoming).
8   JJ Erker, C Goanta , G Spanakis, 'Cancel Culture Corpus Through the Lens of Natural Language Processing' in *Proceedings of the First Workshop on Language Technology and Resources for a Fair, Inclusive, and Safe Society within the 13th Language Resources and Evaluation Conference* (European Language Resources Association 2022).
9   SH Mansur, S Salma, D Awofisayo and K Moran, 'AidUI: Toward automated recognition of dark patterns in user interfaces' in *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)* (IEEE Computer Society Conference Publishing Services 2023).

II. Model Building: Detecting and Classifying Violations

Data collection is needed for measuring the prevalence of phenomena but it can also be used for developing models capable of detecting, classifying, and flagging violations. This involves creating algorithms that can identify specific patterns in the data, from manipulative design features to instances of undisclosed advertising. While explaining the essence of a machine learning model is outside of scope of this editorial, it's worth noting that any model built requires a fair amount of good quality data (probably annotated with some legal framework by relevant experts[10]). At the same time, building these models is an iterative process, requiring constant refinement. Collected data must be updated regularly to account for changes in social media affordances or domain specific trends (such as influencer marketing practices), which brings us to the cornerstone of any automated system, the evaluation.

III. Evaluation: Ensuring Accuracy, Fairness, and Accountability

Evaluation is crucial to ensuring that the ML models employed for consumer forensics are accurate, fair, and transparent. Models must be tested against ground truth data to ensure they correctly identify violations without generating excessive false positives or negatives. Evaluation metrics from the ML toolbox are necessary, however avoiding unintended biases and over-enforcement (esp. for certain communities) is essential. To ensure fairness and transparency in enforcement, machine learning models should not operate as black boxes. Explainable AI (XAI) techniques can provide insights into how decisions are made.

For example, if an influencer's post is flagged for failing to disclose sponsorship, the system should provide an explanation that outlines the key factors behind this decision, such as language used or visual cues in the content.[11] Additionally, forensic data collection methods can maintain audit trails, preserving digital evidence that regulators can use to review enforcement decisions. Accountability is critical, especially when enforcement actions may result in significant fines or restrictions on digital services. By incorporating XAI and forensic audit trails, regulators can ensure that platforms and individuals have recourse to challenge incorrect decisions.

In conclusion, building robust frameworks for consumer forensics requires an intricate balance of data collection, model development and evaluation. The DSA's role in streamlining access to data and enforcing transparency will be a game changer for researchers and regulators alike. With frameworks like explainable AI and forensic audit trails, we are better equipped to navigate the complexities of digital spaces, from detecting manipulative design to addressing effective content moderation. However, these advancements only scratch the surface of what's possible. As the digitization of government services continues, there is also a need for investing further in resources (both human and technological) that will allow authorities to autonomously function when they monitor and enforce regulations.[12] Moreover, a strong need for more interdisciplinary research (involving pubic administration, law scholars, technologists, behavioral and social scientists) is needed for effective regulation of new technologies.[13] In many ways, we are just at the beginning of the journey in consumer forensics—and the road ahead promises to be as innovative as it is essential in order to protect consumers across the ever-evolving digital landscape.

Gerasimos (Jerry) Spanakis*

10 F Zufall, M Hamacher, K Kloppenborg and T Zesch, 'A Legal Approach to Hate Speech – Operationalizing the EU's Legal Framework against the Expression of Hatred as an NLP Task' (2022) In Proceedings of the Natural Legal Language Processing Workshop 2022 (Association for Computational Linguistics 2022).
11 T Bertaglia, S Huber, C Goanta, G Spanakis and A Iamnitchi 'Closing the loop: Testing chatgpt to generate model explanations to improve human labelling of sponsored content on social media' in *Explainable Artificial Intelligence First World Conference, xAI 2023, Lisbon, Portugal, July 26–28, 2023, Proceedings, Part II* (Springer 2023).
12 C Goanta and G Spanakis, 'Discussing The Legitimacy of Digital Market Surveillance' (2022) 2(1) Stanford Computational Antitrust 44.
13 C Goanta, N Aletras, I Chalkidis, S Ranchordás and G Spanakis, 'Regulation and NLP (RegNLP): Taming Large Language Models' in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics 2023).
*   Gerasimos (Jerry) Spanakis is an assistant professor in data mining and machine learning at Maastricht University; email: jerry.spanakis@ maastrichtuniversity.nl.