Improving the Efficiency of Beverage Refill Rates in Restaurants
(Technical Paper)

Social Media Content Moderation: An Ethical Consideration
(STS Paper)

A Thesis Prospectus Submitted to the
Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia
In Partial Fulfillment of the Requirements of the Degree
Bachelor of Science, School of Engineering

Daniel Ayoub
Fall, 2019

Technical Project Team Members
William Define
Adam El Sheikh
James Garcia-Otero
Taylor Kramer

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Signature  Date 05/07/2020
Daniel Ayoub

Approved _____ Date _____
Harry Powell, Department of Electrical and Computer Engineering

Approved _____ Date _____
Thomas Seabrook, Department of Engineering and Society

Introduction:

The Christchurch mosque shootings of March 2019 killed 51 people and injured 49 others. The attack was live-streamed on Facebook, where it was viewed less than 200 times during the live broadcast. “Including the views during the live broadcast, the video was viewed about 4,000 times in total before being removed from Facebook” (Rosen, 2019). For 12 whole minutes after the end of the live stream, the original video was available on the largest social media platform in the world for viewing or distribution. Within the first 24 hours alone, Facebook removed 1.2 million videos of the attack at upload. Millions of additional copies spread like wildfire to other platforms such as Reddit and YouTube, prompting a response from those platforms as well. In fact, Facebook “saw a core community of bad actors working together to continually re-upload edited versions of this video in ways designed to defeat [Facebook’s] detection” (Rosen, 2019). All this prompts the question: What role do content moderation schemes have in situations such as these?

Social media platforms have given individuals a voice of free speech, a voice that can be potentially heard around the world in a matter of seconds. With this voice, many people have brought to light vital initiatives to further the greater well-being of our society. People have used social media to advocate for a particular charity or cause, to give others access to education, to support positive political movements, and to communicate with loved ones who may be far. However, despite all these positive uses, some have used social media platforms as a forum to spread hate speech, encourage violence, and popularize terrorism. In response, social media platforms have rushed to create complex algorithms that detect and block offensive and harmful content. In addition, they have hired thousands of human moderators who are responsible for the manual oversight and training of the algorithms. “More than a hundred thousand people work as

online content moderators, viewing and evaluating the most violent, disturbing, and exploitative content on social media" (Chotiner, 2019). The decisions the human moderators make in conjunction with the algorithms have massive implications, ones that can completely change a society for the better or worse. Alongside the social impact of content moderation, there is an ethical consideration that arises from asking human moderators to watch and flag gruesome and potentially psychologically damaging material. Casey Newton, a journalist for *The Verge* writes, "Collectively, the employees described a workplace that is perpetually teetering on the brink of chaos. It is an environment where workers cope by telling dark jokes about committing suicide, then smoke weed during breaks to numb their emotions" (Newton, 2019).

In the STS paper, I will first report the status quo on current social media content moderation techniques. Then, I will argue that the use of humans as moderators of dangerous content on social media platforms is unethical. Finally, I will suggest alternative techniques to content moderation independent of direct human involvement and evaluate their merit as compared to the status quo.

Social media content moderation is a pivotal technology with many societal and political consequences. However, not all technology falls in this category. In fact, most of the technology we use on a day to day basis has a relatively minimal significance on our society as a whole, yet has considerable impacts on all of our personal lives. My group and I have chosen to focus on a technical invention that improves dining-out customer service satisfaction. We propose a Smart Coaster system to improve the efficiency of refilling drinks in a restaurant. This Smart Coaster system will improve consistency of drink refill rates, and reduce time and effort expended by both servers and patrons. In the technical paper, I will detail the design and development of the invention and the results of our efforts.

Technical Topic:

During a restaurant outing, the most frequent point of interaction between servers and patrons is the process of drink refills. This is a multi-element system that requires the server to repeatedly check each table for near-empty drinks to refill. This puts strain on the customer experience and can potentially reflect negatively upon the establishment. Contemporary hospitality management research has demonstrated that the quality of the restaurant's physical environment and its image are significant predictors of customer perceived value. Customer perceived value is a significant determinant of customer satisfaction, which is a significant predictor of future behavior (Ryu, Lee, & Kim, 2012). Even if the food is excellent, poor customer service can deter patrons from ever coming to the restaurant again. This problem is exacerbated in especially large restaurants and bars where hundreds of patrons are served in one evening. Even with an adept serving staff, attending to everyone's drink can quickly become overwhelming. To remedy this problem, Adam El Sheikh, James Garcia-Otero, Taylor Kramer, William Define, and I propose a Smart Coaster system to improve the consistency of drink refill wait times. The Smart Coaster will use force sensing to determine if drinks placed on it are nearly empty and if so, will wirelessly notify an application running at the central server station. The server would then respond to this notification by going to the empty drink and refilling it, thereby improving the overall service of the restaurant.

The project will involve the design and construction of three main components: the coaster, the induction charging station, and the central server station application. Additionally, the coaster will include WiFi connectivity and wireless charging. These two features allow for seamless and flexible integration in a restaurant setting which potentially may have hundreds of customers at a collection of tables. Although the basis of this project is in electrical and computer

engineering design, it requires understanding and consideration at a system level. This project solves a real-world problem and has the opportunity to positively impact multiple parties.

Similar attempts to address drink refill issues have occurred in academia. The most similar attempt was from students at Saarland University in Germany in 2005 (Butz & Schmitz, 2005). Students designed a beer mat to detect the weight of empty drinks and incorporated coaster-flipping detection to be used in "voting" games. Although the Saarland University project is very similar, the smart coaster proposed in this project incorporates induction charging which sets it apart from competitors and takes into consideration scalability. Additionally, University of Virginia students have previously attempted to solve the problem of automatic drink detection and serving through the use of a robotic arm with computer vision (Hutchinson, Gustafson, Houska, & Syed, 2019). This previous project is similar in purpose to the proposed project, however, it introduces ethical issues and creates opportunities for misinterpretation with varied lighting sources. The proposed smart coaster is scalable, easy-to-integrate, and aims to increase server efficiency without removing the need for a waiter or introducing image security concerns.

STS Topic:

Social media content moderation represents an intricate balance and one that continues to elude many modern-day social media companies due to its immense social, ethical, and human dimensions. Get social media moderation wrong, and hate speech may proliferate through the masses. Moderate too much, and social media limits people's use of free speech and the platform loses its appeal. Moderate with human actors or ineffective technologies, and ethical concerns surface. The legal aspects attributed to this type of content moderation are particularly

interesting. “Unlike public institutions, social platforms are private entities that are under no legal requirement to grant anyone the right to speak or participate in their walled gardens. This affords them the unique ability to both inject themselves as intermediaries between citizens and their governments and decide which voices may be heard” (Leetaru, 2018). In other words, social media platforms have essentially full autonomy in deciding what can and cannot be heard. These platforms are not subject to the legal scrutiny faced by traditional media platforms such as *The Washington Post* and *The New York Times*.

The moderation of social media is undoubtedly a complex topic, one with many actors, stakeholders, and specialized legal status. Examining this topic under the STS framework known as “Sociotechnical Transitions” (also known as the “Multi-Level Perspective on socio-technical transitions”) will help facilitate a greater understanding of social media content moderation with respect to the research questions. “The approach suggests that diffusion or transitions occurs through interactions among three levels: the niche, the regime, and the landscape” (Sovacool & Hess, 2017). The niche refers to a technology or concept that is emerging to viable market introduction or rapidly becoming disruptive. “The regime refers to the incumbent sociotechnical

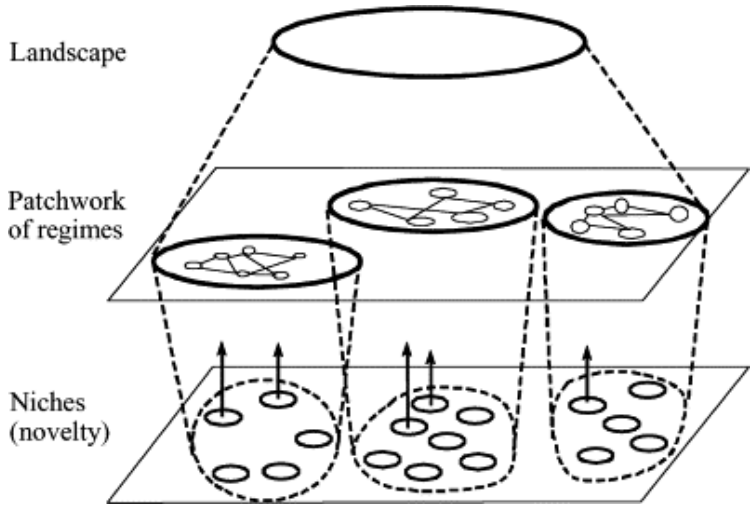

Figure 1: Multiple levels as a nested hierarchy (Geels, 2002)

system that the niche is potentially affecting or replacing; such regimes contain cognitive, regulative, and normative institutions" (Sovacool & Hess, 2017). Finally, the landscape refers to major external developments or disruptions such as wars, changes in policy, economic fluctuations, etc. These external disruptions pressure the regimes to adapt which in turn can create opportunities for the niches to grow or diffuse accordingly. Figure 1 above shows the structure inherent in this framework. "The nested character of these levels, means that regimes are embedded within landscapes and niches within regimes" (Geels, 2002).

The analysis of social media content moderation fits exceptionally well within this framework. For this paper, the specific methodologies used by companies to moderate the content on their platforms will be designated as the niches. Under this definition, the use of computer algorithms and the use of human moderators would be two separate niches. The regime in question is undoubtedly the social media platform. For example, YouTube would be a regime, separate from Facebook, separate from Twitter. Lastly, the landscape is the grander political climate in which the social media is operating. "A significant benefit to the theory is its emphasis on interactivity and dynamic interactions among the three levels of niches, regimes, and landscapes" (Sovacool & Hess, 2017). Therefore, the use of this framework will elucidate the relationship between the current political climate, the social media platform, and the moderation techniques employed. Examining these relationships, one can learn the origin and rationale behind the current content moderation techniques available. For example, using this framework, one can ask if the political landscape imparts pressure on social media companies to hire human moderators. From this, one can glean together details of the external pressure on social media platforms and ultimately the ethics of hiring humans as content moderators. One can also examine the alternative niches (other content moderation techniques) that are emerging into the

system and the responses of both the regime and the greater landscape to the changes. This framework does not guarantee that these new niche-innovations will inevitably win over existing niches (Geels, 2019). This specific framework allows for an analysis of the interplay between the content moderation technique, the social media, and the government or political circumstances surrounding the content. The figure below illustrates the full range of possible transitions between niches, regimes, and landscapes.

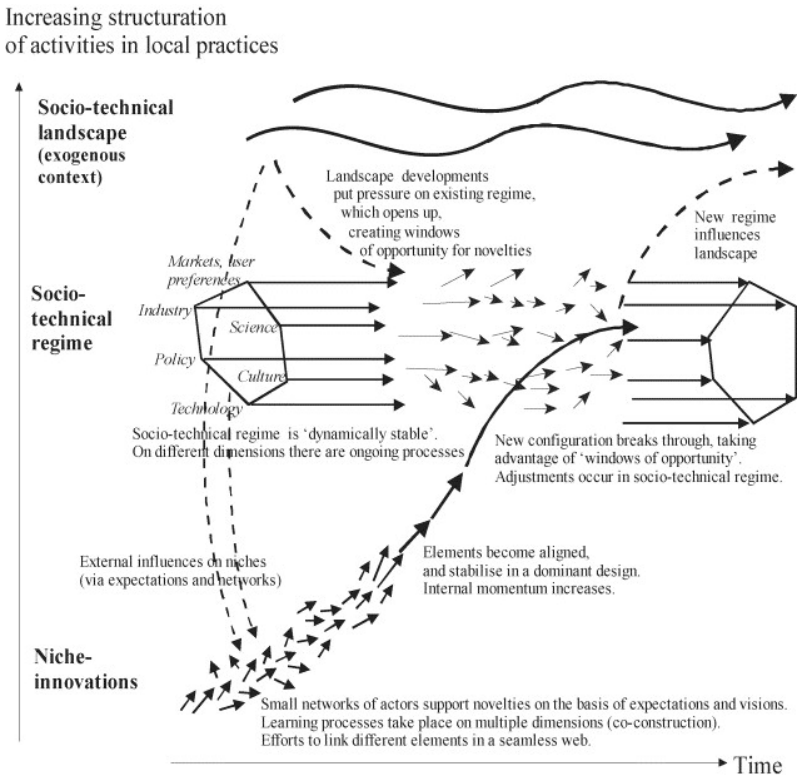

Figure 2: Multi-level perspective on socio-technical transitions (Geels & Schot, 2007)

Research Questions and Methods:

Content moderation is a loaded and ambiguous concept due to the lack of standardization across social media platforms, the inherent subjectivity in determining content guidelines, and the unforeseen consequences that often arise from those decisions. At the Conference of Human Factors in Computing Systems in 2017, researchers discovered "platform definitions of what

constitutes behavior like 'harassment' are incredibly vague and also highly inconsistent across platforms" (Wohn, Fiesler, Hemphill, De Choudhury, & Matias, 2017). At Facebook, for example, employees who meet to set the content guidelines (mostly young engineers and lawyers) "try to distill highly complex issues into simple yes-or-no rules" (Fisher, 2018). These vague guidelines are then distributed to content moderators who have just a few seconds to determine whether a particular post fits the rules. *The New York Times* has found that "the closely held rules are extensive, and they make the company a far more powerful arbiter of global speech than has been publicly recognized or acknowledged by the company itself" (Fisher, 2018).

In an effort to focus this research, I hope to ask and answer the following question: What are the ethical considerations surrounding the employees hired to moderate harmful content? Social media platforms take all the content they want to shield from the greater world, and they ask a smaller subset of humans to watch it and make decisions regarding it. Is this ethical just because a smaller subset of humans is involved? Does it make it any less dangerous? Lastly, I hope to explore alternative techniques to content moderation and evaluate their effectiveness relative to the status quo.

To answer these above questions, I will use a mix of two different methodologies: policy analysis and network analysis. Policy analysis allows for an investigation into the formation and evaluation of organizational, city, state and federal policies. By investigating the moderation policies in place at various social media companies, I can better understand the content moderation status quo. This method can also help me suggest alternatives that still line up with existing greater company policies. Network analysis allows for an investigation into the relationships existent between the government, social media platforms, and the moderation

techniques used. This methodology works exceptionally well within the STS framework outlined in a preceding section.

Conclusion:

At the end of my research, I should have three deliverables: a technical research paper, a technical product implemented to solve the objective, and an STS research paper. The technical product, the Smart Coaster, will improve drink refill times, and thus customer service satisfaction in restaurants. This technical work is an example of how a small technology, while not necessarily pivotal to greater society, can impact and influence lives positively. The technical paper will report on the design, implementation, and test of the Smart Coaster. It will outline the scalability of the device, highlight new innovations, and present the challenges faced. The STS research paper aims to answer if the use of human employees as content moderators for social media platforms is ethical. This paper will also examine alternative techniques to content moderation and the role these new techniques will potentially have. If I am successful in answering both of these concerns, the STS research paper could very well be publishable in a broader STS research journal. If the technical product works well and is delivered according to specification, my group and I could pursue a patent on the idea, as it is a relatively new concept, and there does not exist much prior art.

References:

Butz, A., & Schmitz, M. (2005). *Design and applications of a beer mat for pub interaction*.

Chotiner, I. (2019, July 5). *The Underworld of Online Content Moderation*. Retrieved from https://www.newyorker.com/news/q-and-a/the-underworld-of-online-content-moderation

Fisher, M. (2018, December 27). Inside Facebook's Secret Rulebook for Global Political Speech. *The New York Times*. Retrieved from https://www.nytimes.com/2018/12/27/world/facebook-moderators.html

Geels, F. W. (2002). Technological transitions as evolutionary reconfiguration processes: A multi-level perspective and a case-study. *Research Policy*, *31*(8), 1257–1274. https://doi.org/10.1016/S0048-7333(02)00062-8

Geels, F. W. (2019). Socio-technical transitions to sustainability: A review of criticisms and elaborations of the Multi-Level Perspective. *Current Opinion in Environmental Sustainability*. https://doi.org/10.1016/j.cosust.2019.06.009

Geels, F. W., & Schot, J. (2007). Typology of sociotechnical transition pathways. *Research Policy*, *36*(3), 399–417. https://doi.org/10.1016/j.respol.2007.01.003

Hutchinson, K., Gustafson, J., Houska, B., & Syed, M. (2019). Aquarius Drink Filling Robot. *The Oculus: The Virginia Journal of Undergraduate Research*, *17*. Retrieved from https://issuu.com/theoculus/docs/oculus_2018-19

Leetaru, K. (2018, September 8). Is Social Media Content Moderation An Impossible Task? Retrieved September 22, 2019, from Forbes website: https://www.forbes.com/sites/kalevleetaru/2018/09/08/is-social-media-content-moderation-an-impossible-task/

Newton, C. (2019, February 25). The secret lives of Facebook moderators in America. Retrieved September 22, 2019, from The Verge website: https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona

Rosen, G. (2019, March 20). A Further Update on New Zealand Terrorist Attack | Facebook Newsroom. Retrieved October 16, 2019, from https://newsroom.fb.com/news/2019/03/technical-update-on-new-zealand/

Ryu, K., Lee, H.-R., & Kim, W. (2012). The influence of the quality of the physical environment, food, and service on restaurant image, customer perceived value, customer satisfaction, and behavioral intentions. *International Journal of Contemporary Hospitality Management*, *24*, 200–223. https://doi.org/10.1108/09596111211206141

Sovacool, B. K., & Hess, D. J. (2017). Ordering theories: Typologies and conceptual frameworks for sociotechnical change. *Social Studies of Science*, *47*(5), 703–750. https://doi.org/10.1177/0306312717709363

Wohn, D. Y., Fiesler, C., Hemphill, L., De Choudhury, M., & Matias, J. N. (2017). How to Handle Online Risks?: Discussing Content Curation and Moderation in Social Media. *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems - CHI EA '17*, 1271–1276. https://doi.org/10.1145/3027063.3051141