Trkic G00gle: Why and How Users Game Translation Algorithms

SOOMIN KIM, Seoul National University, South Korea
CHANGHOON OH, Boston College, USA
WON IK CHO, Seoul National University, South Korea
DONGHOON SHIN, Seoul National University, South Korea
BONGWON SUH, Seoul National University, South Korea
JOONHWAN LEE*, Seoul National University, South Korea

Individuals interact with algorithms in various ways. Users even game and circumvent algorithms so as to achieve favorable outcomes. This study aims to come to an understanding of how various stakeholders interact with each other in tricking algorithms, with a focus towards online review communities. We employed a mixed-method approach in order to explore how and why users write machine non-translatable reviews as well as how those encrypted messages are perceived by those receiving them. We found that users are able to find tactics to trick the algorithms in order to avoid censoring, to mitigate interpersonal burden, to protect privacy, and to provide authentic information for enabling the formation of informative review communities. They apply several linguistic and social strategies in this regard. Furthermore, users perceive encrypted messages as both more trustworthy and authentic. Based on these findings, we discuss implications for online review community and content moderation algorithms.

CCS Concepts: • **Human-centered computing** → **User studies**.

Additional Key Words and Phrases: Human-AI Interaction; algorithmic experience; gaming; translation algorithm; online review; recommendation algorithm; peer-to-peer platform

ACM Reference Format:
Soomin Kim, Changhoon Oh, Won Ik Cho, Donghoon Shin, Bongwon Suh, and Joonhwan Lee. 2021. Trkic G00gle: Why and How Users Game Translation Algorithms. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 344 (October 2021), 24 pages. https://doi.org/10.1145/3476085

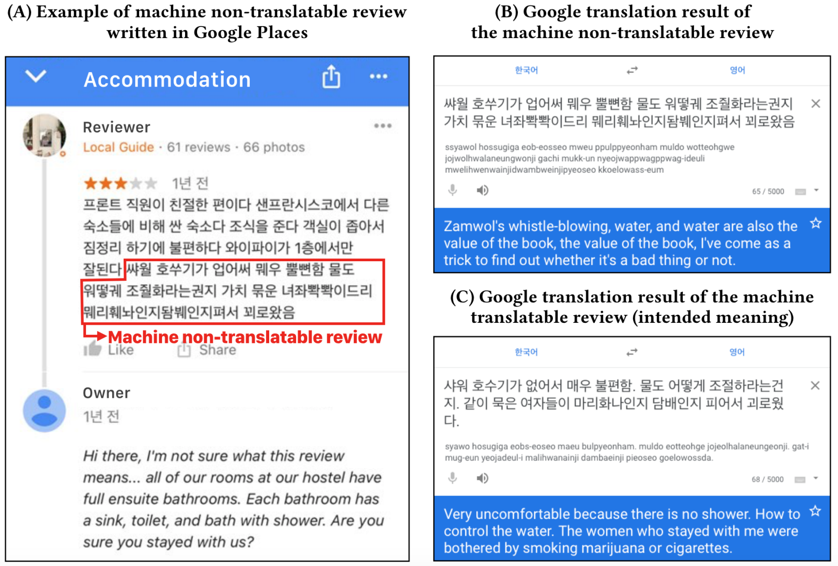
1 INTRODUCTION

People interact with algorithms in various ways as AI has increasingly found its way into our daily lives. YouTube users constantly watch video clips as the curation algorithm suggests, and Amazon consumers often add the products that it suggests. Furthermore, users attempt to actively engage with such algorithms. For example, when users are facing issues with curation or recommendation

[*] Corresponding author

Authors' addresses: Soomin Kim, Seoul National University, Address, City, South Korea, soominkim@snu.ac.kr; Changhoon Oh, Boston College, Address, City, USA, changhoon.oh@bc.edu; Won Ik Cho, Seoul National University, Address, City, South Korea, tsatsuki@snu.ac.kr; Donghoon Shin, Seoul National University, Address, City, South Korea, ssshyhy@snu.ac.kr; Bongwon Suh, Seoul National University, Address, City, South Korea, bongwon@snu.ac.kr; Joonhwan Lee, Seoul National University, Address, City, South Korea, joonhwan@snu.ac.kr.

2573-0142/2021/10-ART344 $15.00
https://doi.org/10.1145/3476085



(A) Example of machine non-translatable review written in Google Places

Reviewer
Local Guide · 61 reviews · 66 photos
★★★☆☆ 1년 전
프론트 직원이 친절한 편이다 샌프란시스코에서 다른 숙소들에 비해 싼 숙소다 조식을 준다 객실이 좁아서 짐정리 하기에 불편하다 와이파이가 1층에서만 잘된다 쌰월 호쑤기가 업어써 뭬우 뿔뼌함 물도 워떻궤 조쥘화라는권지 가치 묶운 녀좌빡빡이드리 뭬리훼놔인지됨뷉인지펴서 꾀로왔음

Owner
1년 전
Hi there, I'm not sure what this review means... all of our rooms at our hostel have full ensuite bathrooms. Each bathroom has a sink, toilet, and bath with shower. Are you sure you stayed with us?

(B) Google translation result of the machine non-translatable review

쌰월 호쑤기가 업어써 뭬우 뿔뼌함 물도 워떻궤 조쥘화라는권지 가치 묶운 녀좌빡빡이드리 뭬리훼놔인지됨뷉인지펴서 꾀로왔음

ssyawol hossugiga eob-eosseo mweu ppulppyeonham muldo wotteohgwe jojwolhwalaneungwonji gachi mukk-un nyeojwappwagppwag-ideuli mwelihwenwainjidwambweinjipyeoseo kkoelowass-eum

Zamwol's whistle-blowing, water, and water are also the value of the book, the value of the book, I've come as a trick to find out whether it's a bad thing or not.

(C) Google translation result of the machine translatable review (intended meaning)

샤워 호수기가 없어서 매우 불편함. 물도 어떻게 조절하라는건지. 같이 묵은 여자들이 마리화나인지 담배인지 피어서 괴로웠다.

syawo hosugiga eobs-eoseo maeu bulpyeonham. muldo eotteohge jojeolhalaneungeonji. gat-i mug-eun yeojadeul-i malihwanainji dambaeinji pieoseo goelowossda.

Very uncomfortable because there is no shower. How to control the water. The women who stayed with me were bothered by smoking marijuana or cigarettes.

Fig. 1. (A) illustrates an example of the machine non-translatable review and the host's response to it as posted in Google Places. The machine non-translatable content is highlighted in red. Both reviews in (B) and (C) include identical content of the highlighted part in (A). (B) presents the result of the translation using Google translate. The reviewer's intended meaning is presented in (C). By encrypting the messages of (C) using a morpho-phonological trick, the reviewer wrote the machine non-translatable review as be seen in (B).

algorithms, they indirectly nudge the algorithm to show their preferred contents by intentionally visiting a certain page more often, or by altering their profile settings [12].

Recent examples have made it clear that users deliberately "trick", "game", or "astroturf" AI algorithms, thereby generating adversarial examples [13, 79]. The concept of "tricking" has been used to refer to a type of user behavior in which the user manipulates the system through the application of several strategies [79]. Similarly, the construct of "gaming" has recently been discussed given a more theoretical background. "Gaming" highlights user behavior with respect to manipulating or deceiving algorithmic systems to fulfill the users' own goals and needs [10, 27, 82]. Since the terms "tricking" and "gaming" have analogous meanings, we use these terms interchangeably throughout the paper, all the while building on prior theoretical research related to gaming behaviors. Völkel et al. [79] found that users can deceive an automatic personality assessment algorithm for the sake of protecting their privacy by utilizing a number of strategies, such as varied language style and word choice. Teenagers are using group accounts in order to trick the Instagram tracking algorithm [65]. In online communities, users deceive censoring and moderation algorithms by deforming their language in an aversive way not interpretable by machines. Chinese netizens constantly devise codewords for freedom of expression in an effort to avoid political and algorithmic censorship [31, 71]. They often employ diverse strategies including using homophonic, logographic and allusory codewords [76]. Similarly, MMORPG users invented a number of codewords to detour the keyword matching algorithm which blocks certain words, including 'freedom', from in-game chat [80]. Furthermore, users circumvent and control natural language processing algorithms by manipulating their own messages in social communities [7].

Users also game translation algorithms. Machine translation algorithms enable communication across language barriers and are applied in a wide range of peer-to-peer platforms, from accommodation platforms (Airbnb), to local business platforms (Google Places) to commerce platforms (Amazon). While translation algorithms are intended to support multilingual communication, recent cases have shown that a number of users write machine non-translatable reviews in order to deceive hosts or business owners (Figure 1). So far, however, there has been insufficient discussion about the why and how users trick the translation algorithms that were initially developed to improve interpersonal communication. By understanding users' underlying motivations along with their strategies, we can acquire a deeper understanding of new human behavior in the algorithmic era, as well as discern insights into how to better design online communities and algorithm-based interfaces.

In order to understand users' behavior of 'gaming/tricking the system', we propose to include diverse stakeholders around the algorithm. By analyzing a broad set of stakeholders around the algorithms, rather than focusing on the sole gaming actor, we elucidate a comprehensive understanding of the social context around the algorithmic system [27].

To this end, this study aims to explore the aspects surrounding interactions in which users trick algorithms, featuring a primary focus on peer-to-peer review communities. We investigated a case in which users write reviews in such a way where the translation algorithm is unable to interpret. We would like to understand how diverse stakeholders interact with and influence each other [10, 27]. Furthermore, we discuss the potential implications of this for future interfaces found in online peer-to-peer communities where user-generated information is created, curated, and consumed. To our knowledge, this is the first study to demonstrate user behavior perpetrated to deceive algorithms in the online review community.

To investigate and understand users' deception of tricking algorithms, we carried out a series of user studies. First, we conducted in-depth interviews to understand users' motivations for gaming algorithms; 14 participants who had experience in writing reviews aimed at deceiving machine translation algorithms were interviewed. Second, we classified users' writing strategies in order to trick translation algorithms based on 87 actual reviews. Finally, to understand readers' perceptions of machine non-translatable, encrypted reviews, we conducted a user study investigating the effects of machine translatability and valence with respect to user perception. The results of this research can be summarized as follows:

- Diverse stakeholders including hosts, reviewers, potential customers, companies/platforms and algorithms are involved in intricate interaction around users' gaming behavior.
- Users trick the translation algorithm in online reviewing communities in order to avoid censorship, to reduce interpersonal burden, to protect privacy, and to provide authentic information with the express goal of creating sincere review communities.
- Users subvert the translation algorithm by following the strategies hereafter: (1) morphological, (2) morpho-phonological, (3) optical, (4) semantic, and (5) mixed tricks.
- Potential customers perceive the encrypted, machine non-translatable review in light of negative information to be less informative, though more trustworthy and authentic.

Based on these findings, we have put forward design considerations for building a better online review community and AI algorithms models capable of involving users in the algorithm development process. The main contributions of this work to the HCI community are as follows:

- We explored why and how users trick algorithms in peer-to-peer review communities in which a broad set of stakeholders interact through both quantitative and qualitative approaches.
- We classified user strategies in deceiving machine translation algorithms, suggesting how this typology can be applied to content moderation algorithms.

- Finally, we discussed implications for designing peer-to-peer review communities in view of the aspects of user interaction with algorithms.

Before elaborating further on our work, we would like to clarify some of the terms frequently used throughout this paper. Taking into consideration the different perspectives surrounding the translation algorithm, we defined the concept of ‘translatability’ from the machine perspective, while ‘encryption’ is taken from the perspective of the human reviewers. Thus, ‘non-translatable reviews’ refer to the reviews that are difficult to algorithmically translate with high accuracy since human users can intentionally ‘encrypt’ messages to deceive a translation algorithm. We use the terms interchangeably depending on the context throughout the rest of the paper. We also use the term ‘stakeholders’ to refer to relevant actors who are interacting with and around algorithms and systems [10, 27]. These terms will be described in more detail in the following sections.

2 RELATED WORK

Our research draws on related work from two areas. First, we review research on how people interact with algorithms, particularly focusing on how they game algorithms. Second, we review works on algorithm mediated communication focusing on the utilization of translation algorithms.

2.1 How Users Interact with and Game Algorithms

Algorithms have become an integral part of everyday life as they are applied in various systems that are utilized on a daily basis. For example, recommendation and ranking algorithms are widely used in reviewing platforms and social media for content curation [7, 15, 17]. Language translation algorithms are used to facilitate interactions between users, regardless of their region and language [54].

Users interact with algorithms not only by accepting their output, but also by directly and/or indirectly affecting them. Inferring how the algorithms work, some users tailor algorithms for their convenience. In particular, users selectively consume specific information to adjust content recommendation and curation algorithms [12]. For instance, YouTube users have been shown to view certain types of video clips to continuously receive recommendations from similar genres. In social media, users have tried to influence the newsfeed by nudging the algorithm to reveal specific content (i.e., visiting certain pages and liking certain posts more often).

Users form certain beliefs and have a mental model toward AI as AI systems are used on a daily basis. People have folk theories on automated curation systems [14]. For instance, users have various beliefs about the Facebook news feed curation algorithm, including passive consumption, producer privacy, consumer preferences, missed posts, violation expectations, and speculation about the algorithm [73]. While users have concerns on the curation algorithm, they apply several coping strategies to overcome the concerns and resolve their violated expectations [73]. In some cases, curation platforms can expose biased reviews. Rather than accepting those reviews without doubt, users discuss the rating system, raise issues about other users’ rating biases, and reverse-engineer the rating algorithm [16].

Moreover, beyond reasoning algorithms and utilizing them for better recommendations, some users even trick or game such algorithms. Research on gaming algorithms is extensive in the field of computer science with a growing base of literature on adversarial machine learning. In this work, we use the metaphor of *game* or *trick* to describe the behavior of users who intentionally manipulate algorithmic models to obtain preferable outcomes [82]. Users trick and manipulate the algorithms to make their behavior more *algorithmically recognizable* [25]. For example, some teenagers using Facebook have pretended to be pregnant or get married in order to feature their posts [72]. Users also play the “visibility game” in Instagram [31], Youtube [3], and Twitter [7] by strategically manipulating algorithmic visibility. On the other hand, users adjust their actions to be *algorithmically unrecognizable* as well. People can create a falsified profile to confuse an automatic personality assessment algorithm [79]. Weibo users algorithmically circumvent censorship by substituting banned terms with homophones [31]. Twitter users also subtweet and use coded messages to avoid attention from content moderation algorithms [7].

Our work builds on previous research done on users' gaming behavior toward algorithmic systems. While recent works have illustrated users' gaming behavior toward content moderation and curation algorithms [3, 7, 31], little attention has been paid to translation algorithms which observe users' social and linguistic strategies. Furthermore, we focus on global peer-to-peer reviewing platforms that have a broad range of interaction between stakeholders. Along with users' social and linguistic strategies, we uncover their underlying motivation to gain a deeper understanding of this emerging behavior in the algorithmic era.

2.2 Algorithm Mediated Communication and Cross-lingual Communication

As algorithms are introduced into social systems based on communication and collaboration, they influence interpersonal relationships. The addition of algorithms to interpersonal communication represents a new paradigm of Artificial Intelligence-Mediated Communication (AI-MC): "interpersonal communication that is not simply transmitted by technology, but modified, augmented, or even generated by a computational agent to achieve communication goals" [29, 38]. Algorithms can improve romantic relationships by enhancing affectionate communication [44] and can facilitate group discussion by boosting group dynamics [43]. Moreover, we are encountering various examples of algorithm-mediated communication in our daily lives including grammar correction, predictive text, machine translation.

Translation algorithms are used to support cross-lingual communication by augmenting original messages [54]. Researchers have investigated how translation algorithms and their interfaces affect the outcomes of interpersonal communication. Users perceive highlighted translation messages to be more clear and less distracting, promoting a collaborative experience [21]. Beliefs that communicators are using machine translated systems significantly increase the chances of a positive collaboration experience [22]. Based on these findings, attempts have been made to improve multilingual communication and collaboration by making translation systems salient [22] and highlighting the critical portions [69]. Accordingly, Lim et al. [55] developed a system that provides emotional, cultural, and contextual annotations along with machine translation results.

Algorithm-mediated communication utilizing translation algorithms is aimed at improving communication or interpersonal discourse [29]. However, unexpected interaction and consequences occur surrounding translation algorithms, and these can, in turn affect online ecosystems that adapt the algorithms. For instance, Temple and Young [77] raised the issue of the objectivity and neutrality of people participating in the translation process, a problem that can affect machine translation results. A framework was also proposed to deal with the challenges of mutual understanding among participants with diverse linguistic and cultural backgrounds [66]. Furthermore, recent examples have shown that normal users of peer-to-peer platforms are deceiving the translation algorithms that were originally intended to support multilingual communication [54], implying that global peer-to-peer platforms do not satisfy all users. Only users native in the written language can read these reviews and this may lead to information asymmetry. By understanding users' gaming behavior toward machine translation algorithms in these communities, we can establish implications on how to design well-functioning online global review communities that embed translation algorithms.

3 STUDY CONTEXT: GAMING IN PEER-TO-PEER REVIEW COMMUNITY

In this section, we describe a case study of users' gaming behaviors toward translation algorithms in online peer-to-peer communities. We also clarify the related stakeholders that interact with the translated reviews.

While online peer-to-peer platforms, such as Airbnb and Google Maps Places, have been applauded as cross-border global communities due to their ease and access of use for a wide range of demographics, recent examples have shown that a number of users are writing reviews in methods to circumvent translation algorithms.

To understand users' gaming behaviors, we explored on how different stakeholders interact and influence each other. We assume five stakeholders: (1) hosts, (2) reviewers, (3) potential customers, (4) algorithms, and (5) company/platform. We assume the algorithms as a stakeholder based on the Actor-network theory. Actor-network theory regards technology as an actor which interacts with human beings rather than as a merely passive entity [50]. The volume of HCI research has adopted this perspective to understand the technology which interacts with users [45, 47]. Similarly, in the context of our study, users interact with algorithms and develop strategies for deceiving and tricking them.

- ***Host***: The host provides the service and interacts with the algorithms, reviewers, and potential customers. They read and manage reviews written in foreign languages with the help of translation algorithms. Moreover, hosts use strategies to appeal to evaluative ranking algorithms by reasoning how they work (e.g., reverse-engineering the search algorithm and adjusting listing price) [40]. Hosts also take action to attract potential customers, giving a close look at existing guests' reviews [40]. Furthermore, because personal interests are involved, hosts are more directly involved in the review moderation process and manually manipulating the review messages by censoring and managing the reviews [59].
- ***Reviewer***: The reviewer is an information provider who writes reviews based on their own experience. Reviews written by reviewers can inform potential customers of the positive and negative aspects of the service and furthermore penalize hosts if they provided poor quality of service [11]. However, recent evidence has shown that reviewers tend to underreport negative information due to attachment to the host and the influence of other reviews [2, 6]. To resolve this challenge, some users devise strategies to avoid censorship and encrypt messages by counterplotting against the translation algorithm.
- ***Potential Customer***: The potential customer is an information consumer who makes decisions based on information posted by the host and the reviews written by the reviewers. Potential customers shall be regarded as primary stakeholders in the review community in light of that they determine the attitudes toward commodities, as well as purchasing behaviors, based on user reviews and ratings [19, 46]. They base their actions on reduced risk and a perceived trust of the host and service [61]. However, in an online environment in which fake reviews are prevalent, there are challenges in acquiring useful and authentic information that would help in making purchase decisions [57]. Potential customers who consume information written by reviewers may also be related to the reviewers' gaming behavior. Furthermore, it is uncertain how they perceive reviewers' encrypted messages and what factors influence the perception of these reviews.
- ***Algorithm***: Two main algorithms work in online review platforms: a review curation/ranking algorithm and a content translation algorithm. The ranking algorithm sorts reviews in an order that is helpful to users, while the translation algorithm supports users and hosts to communicate and exchange information regardless of nationality and language [11]. This study focuses on translation algorithms. Translation algorithms are employed to augment

messages with the aim of supporting multilingual communication [54]. Since the translation algorithm mediates multilingual interaction, the translation performance is able to affect hosts, reviewers, and potential customers. We attempt to ascertain how various stakeholders relate to translation algorithms and how the translation algorithm affects the actors' behaviors.

- ***Company/Platform***: From the perspective of the company operating the platform, it is important to manage and moderate user-generated content such as reviews in order to manage the brand reputation and promote continuous user participation. However, most companies run an impenetrable review system for several reasons, including preventing deceitful users who play the system and game the norms [74]. These veiled mechanisms could lead to complex interaction patterns and unintended consequences beyond the company's initial planned intent. Although we obviously cannot understand the algorithms and mechanisms intended by the company, we could reflect and reflect the interpersonal and social consequences caused by the platform by exploring the users' interactions around the system.

Among the stakeholders, we focused on the 'reviewers' who deceive a translation algorithm and the reviews that they wrote. Furthermore, we intended to develop a comprehensive understanding of the phenomena surrounding the translation algorithm by interpreting the relationships between the various stakeholders that can be inferred by our study's results. We intended to adopted an analytic perspective of 'playing the game' rather than a 'gaming the system' as proposed by Cotter [10]. Previous research that investigates users' gaming behavior puts a sole emphasis on the interaction *with* the algorithm model. This 'who gamed the algorithm' narrative provides a reductive view that focuses narrowly on 'gamer' rather than encompassing a complex set of actors who interact and influence each other's actions [78]. Algorithmic systems are deployed in a complex social context where multiple stakeholders are involved and influence each other. Thus, we argue that we should consider diverse stakeholders *around* the algorithm [10, 27]. In accordance with this view, we would like to address the following research questions:

- **RQ1**: Why do reviewers write encrypted reviews to trick the translation algorithms?
- **RQ2**: What strategies do reviewers use to trick the translation algorithm?
- **RQ3**: How do users (potential customers) perceive and decipher the encrypted, machine non-translatable reviews?

4 STUDY 1. WHY DO USERS WRITE MACHINE NON-TRANSLATABLE REVIEWS?

To address the research questions, we designed a series of user studies. In Study 1, we tried to investigate reviewers' underlying motivations for tricking and deceiving the translation algorithm. Various stakeholders can influence and interact beyond the user's gaming behavior.

4.1 Method

We tried to recruit people who had an experience of writing reviews in peer-to-peer online communities in a machine non-translatable manner. Since the concept of 'machine translatability' presupposes the user's intention to bypass the algorithm, we recruited users who have experience of intentionally writing encrypted reviews in global peer-to-peer communities. Participants were recruited through public social media posts, posts to online travel-related communities, word of mouth, and personal contacts of the authors. We asked the possible participants to capture and send related reviews in advance by email. We recruited 14 participants (8 female, 6 male), aged 21-35 (M=28.4, SD = 2.89). The participants showed a range of nationalities (6 Koreans, 4 Chinese, 2 Frenches, 2 Japanese). We interviewed participants at a public place they chose. For the participants who were unable to meet offline, the interviews were conducted with virtual calls such as Skype. All of the interviews were recorded and transcribed.

To understand users' motivations for writing encrypted reviews, we conducted in-depth semi-structured interviews. The interviews were mainly focused on participants' motivations for writing reviews by tricking algorithms on peer-to-peer communities. We asked them to access their online accounts of the peer-to-peer community (Airbnb or Google Maps Place) and to share their reviews written in a way that cannot be translated by algorithms. The reviews were used as research probes to facilitate the interview. We encouraged the participants to recall the moment they wrote those reviews and asked them to describe their experiences and why they wrote the reviews in that manner. Several follow-up questions were also asked for participants to elaborate on their thoughts, experiences, and behaviors.

Based on the in-depth interview, we tried to elicit major motivations with a thematic analysis [5]. Three researchers read all the 14 interview scripts from the recorded audio materials until becoming familiar with the data. Then, we iteratively read the recorded data to identify meaningful keywords that indicate the motivation and generated initial codes. This process was repeated several times until the identified themes were saturated. Finally, the themes were defined from the clustered keywords and labeled.

4.2 Results

Through the analysis, we identified four categories of users' motivations for writing machine non-translatable reviews: (1) to avoid censoring, (2) to reduce the relational burden, (3) to manage reputation and to protect privacy, and (4) to provide authentic information to other users.

4.2.1 To avoid censoring. Users wrote reviews by bypassing the translation algorithm to avoid censorship by the hosts. They worried that the host could flag and delete their negative reviews. Participants mentioned that they "I expected the host to delete scratching reviews," so [they] "left a review in a way that couldn't be translated," (P3). Another participant said, "with the development of the translation system, foreign hosts can read and delete negative reviews" (P10). This concern was based on real experience, as noted by P12: "I wrote negative reviews of an Airbnb German accommodation, but they were erased and not exposed. Twice or three times. So since then, I've been writing my reviews in a way that the translator cannot translate." P6 also cited avoiding censorship as a primary motivation, and mentioned specific strategies: "I just put a compliment in the first half, and then wrote an encrypted review in the second half. Then the business hosts think it's a compliment and they don't delete it." The surface-level motivation of all users who write machine non-translatable reviews stems from concerns that negative reviews based on objective facts can be censored and deleted by the host. This perception is related to prior studies indicating that users trick algorithms in social media to avoid political censoring by online moderation algorithms [31]. In the peer-to-peer review communities, users try to expose and avoid the lemon market, where uncertainty about service exists [59, 63]. Furthermore, this interaction pattern implies that the interests of various stakeholders are entangled behind the users' tricking behavior.

4.2.2 To reduce the relational burden. Peer-to-peer platforms such as Airbnb and Google Places are based on an intermediary between individual parties. Users actively used the translation algorithm for fluent communication with the foreign host but also tricked the algorithm at the same time. The motivation for users to write encrypted reviews was to reduce the relational burden caused by the host reading negative reviews. P2 said "because I have interacted with the host offline, I am sorry to write the bad comments." P7 stated that he is assessing not just the hosts, but also their assets, which determine their livelihood: "I essentially rate the hosts as much as I rate their property. I'm afraid that my inconvenience will affect the host's whole livelihood. So I give high ratings, but I encrypt drawbacks so that only limited users can recognize them." This is related to the fact that social influences, such as individual attachments, cause users to misreport review scores [2].

This finding implies the possibility that the algorithm can mediate interpersonal relationships in peer-to-peer communities. P1 mentioned the need of the moderation algorithm to implicitly interpret and deliver the content: “If the algorithm modifies the messages so that the host can view them at a higher-level, rather than seeing my original harsh reviews, the burden of writing an honest review will be diminished.”

4.2.3 To manage reputation and to protect privacy. Managing reputation and protecting privacy emerged as the motivation to write encrypted reviews. Some Airbnb reviewers were hesitant to write a negative review for fear of repercussions in Airbnb’s mutual evaluation system. P11 noted that “I can’t honestly write the review because other hosts will not accept me if my evaluation score is low.” Similarly, P3 also mentioned that “the hosts often penalize me when I leave a bad review. If I write a scratching review, future hosts may not accept me.” This result implies that strategic review behavior can be particularly important in the context of a two-sided reputation system. The fact that the host could know the guest’s personal information also raised users’ concerns: “I encrypted in order not to be damaged because the host can know my contact and personal information” (P7). A P14 said, “I experienced a host complained by calling me directly without politely asking for a deletion through Airbnb. It is burdensome that the host can directly contact me after the service.” These statements correspond with the prior result indicating that people circumvent privacy-eliminating surveillance algorithms by preserving personal autonomy [7, 79].

4.2.4 To provide authentic information for other users and to build a trustworthy review community. The communal motivation of sharing information with other users and creating an informative, genuine community makes users circumvent the translation algorithm. Reviewers tried to prevent other users from experiencing the same inconvenience. P15 mentioned that “I want to give other users helpful information and somehow prevent secondary victims,” and P2 said, “My hope is to reduce the number of victims in the future. Nationality also strengthened the users’ commitment to ingroup support.” Some users mentioned that they did not want people with the same nationality to suffer from similar similar experience: “I was afraid that our country people would be harmed in the foreign place,” (P1) and “It was sincere consideration for my people” (P10). This finding is related to the prior research that found users write reviews because of concerns for other users [81]. A number of moderation algorithms are designed for the purpose of surveilling malicious and harmful online content. However, our results imply that users circumvent the moderation algorithms for the good intention of helping other community members and building trustworthy communities.

5 STUDY 2. WHAT STRATEGIES DO REVIEWERS USE TO TRICK THE TRANSLATION ALGORITHM?

Study 2 was conducted to explore the strategies used to trick and circumvent machine translation algorithms. In this study, we focused on users’ interactions with the algorithmic model. This process allowed us to infer users’ heuristics concerning how translation and moderation algorithms operate and how they circumvent these algorithms.

5.1 Method

Avoiding censorship by making content algorithmically unrecognizable and indirectly conveying true intentions occurs in various cultures [7, 31, 71, 76]. Although we observed the behavior of users from various countries in Study 1, we decided to limit our scope to the reviews written in Korean (Hangul, the Korean alphabet). Our decision to limit our research was based on the realistic condition that researchers should understand all of the different linguistic characteristics to analyze strategies used in different languages. We tried to generalize our analysis by allocating multi-lingual examples into our typology. The multi-lingual examples were based on the literature and examples gathered from Study 1.

We collected encrypted reviews from three data sources: reviews participants in Study 1 wrote, manual browsing, and public recruitment. We gathered 15 Korean reviews from participants in Study 1 (participants in Study 1). In addition, by browsing reviews on Airbnb and Google Maps, we collected 43 machine non-translatable reviews (manual browsing). Two researchers looked through reviews of the accommodations in the five most visited cities in the U.S. including LA, New York City, Honolulu, San Francisco, and Seattle. We collected 50 reviews that native speakers might not be able to interpret with translation algorithms. We manually tested whether Google translation can translate them with high accuracy and only 43 reviews that at least two researchers evaluated as highly accurate was selected. We also invited the public to share reviews written in machine non-translatable ways for Korean online travel community (public recruitment). A total of 87 reviews were collected for our analysis. To analyze patterns of tricking translation algorithms, two researchers iteratively aggregated highly related tricking strategies based on the following linguistic properties: morpho-phonology, writing system, and sociolinguistics.

5.2 Linguistic Characteristics of Hangul

We briefly describe the characteristics of Korean, the language used in our data analysis. Korean uses Hangul, which is one of the well-known featural writing systems. The syllable structure of Hangul is CV(C) and consists of a maximum of three components (two consonants and one vowel), which plays a similar role to the bushu of kanji, but each one represents phonetic properties, similar to the Latin alphabet, rather than an ideographic alphabet. These are referred to as first, second, and third sounds, and there are 19, 21, and 28 candidates, respectively. They make up about 11K characters, of which 2,500 are used in real life, and the morpho-syllabic blocks they form denote syllables.

5.3 Results

We identified five social-linguistic strategies that are applied by users to trick translation algorithms: (1) morphological, (2) morpho-phonological, (3) optical, (4) semantics, and (5) mixed tricks. We did not aimed to cover the exhaustive writing strategies. We instead illustrated the broad range of possibilities on how users trick translation algorithms.

5.3.1 Morphological Tricks. The first strategy is morphological modification of lexicons. The linguistic morphology modification observed here may not parallel general morphological theory. However, when dealing with the words or morphemes that are the fundamental units of a sentence, we referred to their modification as morphology. We observed the decomposition of words or morphemes to the character or sub-character level and also cryptic manipulation.

Jumbled Characters Within/Between Words. Users jumbled characters or sub-characters that make up words within the word or between adjacent words, also known as ‘Spoonerism’ [30]. While machine translation algorithms can not interpret jumbled words, native speakers can swap back the sub-units that compose the word or morpheme and easily recognize the original word [20, 70]. As seen in Table 1, the negative lexicon is decomposed (e.g., from “개쓰레기” to “개”,“쓰”, “레”, “기”) and made up as the swapped output (“개레쓰기”). Similarly, one type of French neologism, Verlan is made by flipping the syllables (from “méchant” to “chantmé”) [24, 52].

Split words and jumble (negative words dispersed within the positive ones). This strategy is an advanced version of character jumbling, which is mainly observed in Korean review gaming. A word can be split into the chunk of the characters (from "개쓰레기" to “개쓰”, “레기”). Users

Strategy	English	Korean	Multilingual
Morphological			
Jumble characters	Trkic Goolge	진짜 개레쓰기	hôte chantmé (hôte méchant)
Split and jumble	Goo Trick Gle	개쓰 진짜♥ 레기 좋아요	
Code-switching	Tリック Gooグル	very 개ssu good 레gi	熱情hulk (熱情好客)
Morpho-phonological			
Phonological similarity	Teulig Gugeul	궥쑬액이	米兔 /RiceBunny (Me Too) 184 (いやよ)
Glottalization		깨쓰레끼	diZZZgustting (disgusting)
Optical			
Character substitution	тяıк G00g|e	ㄱH 스스 근 ㅔ ㄱi	目田 (自由)
Redundant consonants		갰씂렚깂	
Semantic			
Metaphor and sarcasm	Hundo P	Turn on *notepad.exe*	MJK (マジか)
Contrasting homophones		진짜 방 the love 요.	無可phone告 (無可奉告)
Mixed			
Mixed strategies		읍읍♥ 쓔레깂 Very good!	

Table 1. Examples of strategies used to trick translation algorithms

then alternate the parts in between two positive connotation words ("개쓰(negative)", "진짜♥ (positive)", "레기 (negative)", "좋아요 (positive)"). Because the parts of words are distributed in a positive context, the review is likely to recognized as positive information to foreigners. However, the native speaker can recognize the reviewers' intended, encrypted message [26].

Code-switching or romanization (along with positive text). This is a code-mixed version of the above methodology where some of the split words are written in different languages. This can make automatic translation more challenging. Code-switching is occurring in a variety of cultures. For instance, Chinese users criticized hosts for treating guests with disregard using the "熱情hulk". "熱情好客", the original meaning of welcoming guests, has been transformed into "熱情hulk" as a meaning of treating guests like a hulk.

5.3.2 Morpho-phonological Tricks. The second typology is to manipulate words and morphemes based on their phonological characteristics. This trick can also involve morphological changes, depending on the characteristics of the writing system [28]. This means that phonological modification of speech causes a morphological change textually, and accordingly, users can infer the original text to some extent just by sounding out the words.

Phonological Similarity (Phoneme Sequence). This strategy is to use homophones with completely different meanings to deceive algorithms. Unlike algorithms, humans can easily understand homophonic substitutes [31]. Homophones or pseudohomophones can be recognized more easily with some semantic cues that reflect the social context of their use [9]. Chinese "Me Too" movement evades political and algorithmic censoring with "#Rice(米)Bunny(兔)" which is pronounced "mi tu" in Mandarin. In this regard, all the Chinese participants in Study 1 mentioned that they used homophones to make negative reviews algorithmically untranslatable.

Glottalization. Users also apply phonological tricks by glottalizing the consonants of the morpho-syllabic block found in the Korean writing system. This occurs in concurrence with the above homophonic codewords, with the difference being that the resulting pronunciation can be far different from that of the original word.

5.3.3 Optical Tricks. The third trick is done by utilizing vision-related characteristics of the writing system. Unlike the previous approach that utilizes phonetic and phonological similarity, orthographical or allocational properties of (sub-)characters are exploited.

Leetspeak (Character Substitution). Leetspeak refers to the strategy taking advantage of optical similarities and conveying intended meanings with lookalike characters or symbols [4]. Examples include: replacing “i” with “!” or “a” with “@” in the Latin alphabet, and inserting ideographs of similar shape in the Chinese alphabet.

Adding redundant consonants. This strategy is mainly observed in Hangul due to its final consonant system. In Hangul, the last sound consonant can be artificially added so as to make the sequence irrelevant to the message. While the message circumvent machine translation algorithms, it is still readable to native speakers.

5.3.4 Semantic Tricks. As a final strategy, users may contextually and allusively convey messages. To understand semantic tricks, people need historical, cultural, and social background knowledge.

Metaphor and sarcasm (cultural). Figurative languages interpreted with socio-cultural contexts (e.g., sarcasm, ironies, paradoxes, and puns) are used to generate encrypted messages. For example, “turn on *notepad.exe*” is internet slang created by Korean netizens to denote negative content. This expression originates from the fact that an internet user can be sued for writing bad comments on celebrities or public figures in the public space. Users who are familiar with this cultural context understand that “*notepad.exe*" means "I want to write a bad review, but I won’t write it."

Contrasting Similar Homophones in Code-switched Manner. There are homophones that are separately categorized because of their code-switched manner of implementation. For example, *‘the love’* is a positive expression in English, but it means “dirty (deoleob)” in the Korean pronunciation.

5.3.5 Mixed Tricks. A mixed approach denotes any gaming review that incorporates more than one of the above. It includes cases for which the text is translatable only for positive sentences, while not being translatable for the negative sentences. Also, because various symbols are available, contrasting the negative text with positive emojis is possible.

5.4 Hosts’ Reaction to Encrypted Reviews

Among the 87 encrypted reviews, 70 reviews contained both positive and negative information and 61 reviews included host’s comments. While 25 comments were positive acknowledgement (e.g., "I am so happy you like my place.”), 23 comments were reactions that the content of the review was hard to understand (e.g., "I’m not sure what this review means..”). Among the encrypted messages, hosts could interpret only those containing positive information, while they had difficulty understanding the hidden meaning of other negative messages.

From the company’s perspective, the service quality could be improved by providing hosts with information regarding indecipherable, negative reviews. The company should isolate review texts which suffer from poor translation accuracy, and then interpret them through methods other than machine translation (e.g., translation by users, employees and crowdworkers). In addition, the company could deliver information to hosts more effectively by separating negative and positive content contained in the review.

5.5 Summary

We classified the strategies utilized to trick translation algorithms. Users masterfully blended these strategies to generate encrypted messages. Most of the users wrote both positive and negative content and encrypted only negative messages; users consistently distracted hosts by including positive content that hosts could understand. To summarize, we identified why and how users trick translation algorithms. In addition, we found that foreign hosts ignore encrypted negative reviews or have difficulty understanding their meaning. Then, how do users perceive the encrypted reviews? To investigate this research question, we conducted study 3.

6 STUDY 3. HOW DO READERS PERCEIVE ENCRYPTED REVIEWS?

Study 3 aims to investigate how users (potential customers) perceive encrypted messages. Would machine non-translatable reviews actually be of service to receivers? In order to investigate receiver's perception of machine non-translatable reviews, an experiment was conducted focusing on a variable referred to as "machine translatability" throughout Study 3. We defined translatability from the perspective of the machine. Thus, "non-translatable reviews" are those which the algorithms find difficult to translate with high accuracy. From the point of view of the user, non-translatable reviews refer to reviews wherein users intentionally use encryption so that the machine translation system cannot interpret it; users encrypt messages in a way that only certain recipients can interpret [23]. In addition, we examined message valence based on the result where this property influences review trustworthiness [18].

6.1 Method

A scenario-based online experiment was conducted to examine how receivers perceive encrypted reviews with respect to informativeness, trustworthiness, and authenticity. Those variables are significant where it concerns evaluating user experiences of online peer-to-peer platforms [61]. This study used a 2 (machine translatable vs. non-translatable) × 3 (positive vs. negative vs. positive and negative) between-subjects design (N = 180). Participants were exposed to an imaginary scenario, on in which they found themselves searching for an Airbnb location to stay at for a vacation, and a review for a specific accommodation was presented. The reviews used as stimuli differed according to the experimental conditions in which participants were designated (Table 2). Participants were then asked to report on the perceived informativeness, trustworthiness, and authenticity of the reviews in question. Since we also aimed at identifying factors associated with users' perceptions of machine non-translatable reviews, we queried them with open-ended questions as to the reason for the participants' quantitative evaluation.

6.1.1 Participants. A total of 180 participants were recruited from online research panels by a third-party provider in Korea. The study subjects consisted of men and women in their 20s to 40s, considering that 58% of Airbnb's hosts and booking guests worldwide were millennials [37]. It should be noted that it is difficult to verify the demographic information of users who write and consume the machine non-translatable reviews. Although we can infer that our target users are the generations accustomed to transforming and creating online texts, we cannot exclude the possibility of potential bias resulting from the sample of the participants. Interventions and further research should pay attention to factors such as the demographic information of the participants.

To ensure adequate representation of the primary target group, quotas were set by sex, age, and geographic location. All participants have experience of using Airbnb. This qualification was made to partially control for the prior experience of the online review community. Among the participants, 50% were females, and 50% were males, and the average age was 29.89 (SD = 6.39, range = 21-43).

Translatability	Valence (3 conditions)		
	Positive	*Negative*	*Pos+Neg*
Translatable	가격은 저렴하고 위치가 좋습니다. 관광지를 걸어서 다닐 수 있습니다.	벌레가 나오고 엘리베이터가 작습니다. 인종차별 심하고 불친절합니다.	Positive + Negative
(Translation)	The price is cheap and the location is good. You can walk around the tourist attractions.	The bug comes out and the elevator is small. Racism is severe and unkind.	
Non-translatable	가교근 저하렴고 윏칣갋 좋니습다. 관광질흘 걼얿섧 다뉠 쑤 잇씁니다.	벌례가 낣옯곫 옐릐베의터가 작습늬다. 읜종차볼 싀마고 불췬절합늬다.	Positive + Negative
(Translation)	It is good to have a lowering of the bridge muscle. Sightseeing days, it's so bitter.	Beolrye-ga is the site of the Yeolbeom site. It's a wide, wide, and battered joint.	

Table 2. Review messages used in the experiment. Translatable and non-translatable messages contain identical meaning. The results translated by Google Translator are presented in "Translation".

6.1.2 Experimental Stimulus. Participants were exposed to reviews uploaded to a fictional Airbnb accommodation. While the content of the review was identical, the stimulus messages were written in a different form depending on machine interpretability. The review of the machine translatable condition was written in formal sentences that were normally translated on Google Translator; The review of the machine non-translatable condition was in sentences that people of the same language sphere can understand and not be translated by Google translation algorithm. We applied morphological, phonological, and semantic tricks to generate the machine non-translatable message.

In order to manipulate the valence, the advantages of the accommodation (price and location) were included for the positive review, and the disadvantages of the accommodation (cleanliness and host's racism) were described in the negative review. In the positive and negative review condition, information of the pros and cons were included.

6.1.3 Manipulation check for experimental stimulus. In advance of the experiment, we aimed to verify whether users can understand the encrypted messages. Twenty participants were asked to translate the encrypted message to the formal and standard way. All 20 participants read and interpreted encrypted messages without difficulty. They were also asked to rate the valence of the positive and negative messages on 10-point differential scales: positive/negative. A paired samples t-test significantly supports the manipulation of message valence ($t(20) = 4.12$, $p < 0.01$). No statistically significant difference was found between negative and positive messages based on the median (5.5) of the Likert scale (neg: 3.41 vs. pos: 7.52), implying that the valence of the stimulus is unbiased. Furthermore, we also controlled the message length.

6.1.4 Measures. **Informativeness** refers to how useful and helpful a review is and affects review quality [62]. To evaluate perceived informativeness, participants assess the review message on a seven-point Likert scale with the items regarding information and usefulness. **Authenticity** of the message implies that the review is perceived as genuine and real [51]. Perceived authenticity was measured via two questionnaires. Respondents rated the degree to which they agreed with each statement on a seven-point Likert scale: (1) The review is authentic; (2) The reviewer reveals his or her genuine experience [51]. **Trustworthiness** refers to the degree to which users believe and trust the content of online reviews, playing an important role in message acceptance and decision-making process [68]. Two seven-point Likert scales were used to measure perceived trustworthiness of the review: (1) The review is trustworthy, (2) The review is reliable [68]. We also asked **open-ended questions** about the rationale for their perception. In doing so, we elicited social and psychological factors that are related to the perception of machine non-translatable reviews.

6.1.5 Analysis. In order to test whether the main effect and interaction effect exist, we used a factorial ANOVA. ANOVA assumptions of normality and homoscedasticity were verified preceding all statistical analyses. Normality of data was verified by Shapiro Wilk test, and all our data fit significantly to a normal distribution. Brown-Forsythe test revealed that all variables did not show significant differences in variance, satisfying homogeneity assumption.

For the qualitative data, we conducted thematic analysis based on affinity diagramming to cluster high-level concepts and to discover recurrent ideas and themes [5]. Two researchers conducted this thematic analysis. The results of this process provided a deeper understanding of the participants' perception toward encrypted reviews and algorithms.

6.2 Results

6.2.1 Perceived Informativeness. The 2 × 3 ANOVA for perceived informativeness yielded main effect for machine interpretability (F(2, 174)=12.01, p<0.001) and for valence (F(2, 174)=16.19, p<0.001). This result means that the participants perceived the machine translatable reviews more informative than machine non-translatable ones. The post-hoc analysis (Tukey HSD) revealed that significant differences in perceived informativeness according to machine interpretability occurred only when the review contained both positive and negative information(p<0.001). In terms of information valence, users perceived the negative reviews more informative than the positive reviews (p=0.028), and the positive+negative reviews more informative than negative reviews (p=0.003). The ANOVA reveals a significant interaction between machine interpretability and valence (F(2, 174)=20.53, p=0.043). This result implies that the degree to which the valence of information affects informativeness is greater for the machine non-translatable reviews.

6.2.2 Perceived Authenticity. The factorial ANOVA revealed that machine interpretability (F(2, 174)=64.71, p<0.001) and valence (F(2, 174)=23.67, p<0.001) have main effects on perceived authenticity. This effect of perceived authenticity was observed for negative (p<0.001) and positive and negative (p<0.001) messages. Machine interpretability and valence had a significant interaction effect on perceived authenticity (F(2, 174)=4.99, p=0.008). The influence of machine translatability on authenticity was observed for the reviews with negative information. Moreover, the valence has a more significant effect on the machine non-translatable reviews.

6.2.3 Perceived Trustworthiness. Analysis of perceived trustworthiness revealed that there are significant main effects for machine interpretability (F(2, 174)=48.95, p<0.001) and valence (F(2, 174)=29.48, p<0.001). This difference was found for the negative review (p<0.001) and the positive and negative review (p<0.001). A significant interaction between the machine interpretability and the valence was also observed (F(2, 174)=20.53, p<0.001). This means that valence differentially affects the perceived trustworthiness of machine translatable and machine non-translatable reviews. Valence had a greater impact on machine non-translatable reviews. Overall, participants evaluated the machine non-interpretable reviews as more trustworthy. However, this effect was observed only when negative information is included in the message. This result implies that not only the exterior feature of the reviews, but also the context the reviews are written within and the information they contain influence trustworthiness as well as authenticity.

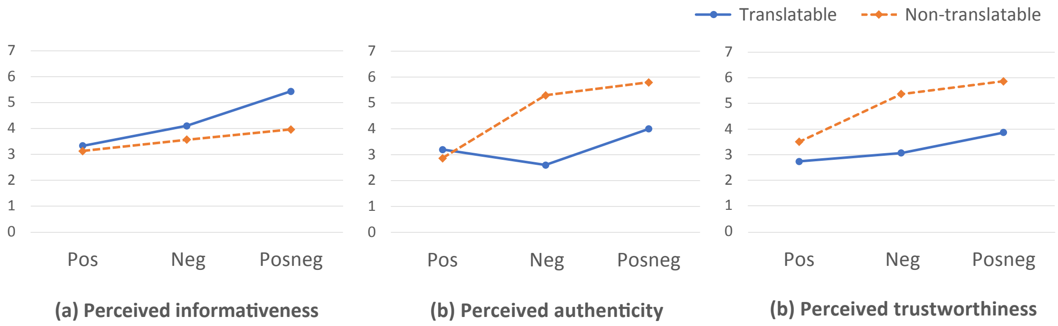


Fig. 2. Experiment result of user perception on reviews. X-axis indicates the valence of the messages. Participants perceived the encrypted reviews as more trustworthy and authentic when they involve negative information. However, the encrypted reviews showed a lower level of informativeness.

6.2.4 *Qualitative Results.*

The results of the qualitative study showed the social and psychological factors that are associated with users' perceptions of the encrypted reviews.

- ***Poor Readability Hinders Informativeness***
  Participants generally evaluated machine non-translatable reviews as less informative due to their poor readability. Although native speakers of the presenting language were able to interpret the encrypted review, the task required a certain task load. A number of participants commented that "This is uncomfortable to read due to typos," and, "It is difficult to understand." In terms of information valence, several users pointed out an issue common to review platforms where exceedingly positive reviews are prevalent: "It is difficult to discern whether the accommodation or service is actually good because there are only good comments. Reviews with pros and cons provide me useful information which is more-so useful."
- ***Machine Non-translatable Reviews Involve Reviewers' Effort***
  Users expressed a perception of machine non-translatable reviews as being more reliable and faithful since they acknowledge the reviewers' effort. Participants noted that the review would likely have been written based on facts due to the writer's endeavor: "If he or she writes with that much effort, I can trust them," and "I feel as though the reviewer wrote a real review while taking hardship."
- ***Reviewer is Good Samaritan with Good Incentives***
  The reviewers' 'good motivation' is what makes machine non-translatable reviews more decent and authentic. Participants noted that they felt the reviewer's consideration for the prevention of other users from experiencing any inconvenience. They mentioned "I feel the review is quite genuine because they don't want others to have the same unpleasant experiences as him or herself," "Kind and gentle.. I am impressed with the reviewer's thoughtfulness not to cause any suffering to in other countries," and "I trust the review because I feel like he or she wants to help out before other victims fall into the same traps." Participants are aware of the reviewer's intentions beyond their message.
- ***Guests Have a Sense of Groupness against Host and Algorithm***
  Users formed a kind of in-group connection with reviewers in opposition of hosts which causes their reviews to appear to be more reliable. This corresponds with the fact that individuals do generally trust smaller, more private, homogeneous groups [60]. Participants drew the line between guests and host: "The review increases the bond between users by writing so that the host and company are unable to understand." The fact that reviewers

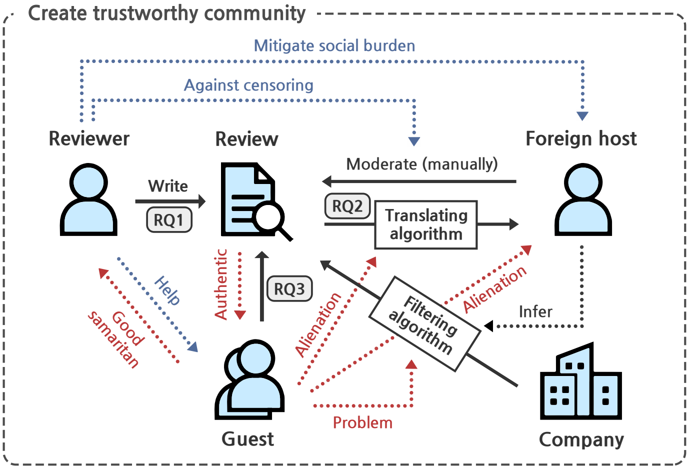


Fig. 3. Schematic representation of the interaction patterns appearing in the global peer-to-peer review community. Various stakeholders interact around the algorithm in the peer-to-peer review platform. The black lines illustrate observable behaviors. The blue dotted lines present the reviewer's motivation (Study 1) and the red dotted lines present guests' perceptions (Study 3). The black dotted line represents the motivations of both reviewers and guests to create trustworthy online communication (Study 1 and Study 3). By analyzing a broad set of stakeholders around the algorithms, rather than focusing on the sole gaming actor, we elucidate a comprehensive understanding of the social context around the algorithmic system.

target a group of the same nationality increased confidence in their reviews. They mentioned thoughts such as "I notice a sincerity wherein the reviewer actually cares about Korean users." Furthermore, participants alienated and otherized algorithms which failed to interpret the encrypted message [67]. One participant compared the AlphaGo and Google translation algorithms: "Google, no matter how many times you have defeated Lee Sae-dol, I think this is quite difficult."

7 DISCUSSION

In this section, we discuss lessons learned from the user studies and implications. We also report our plans for future work as well as the limitations of the study.

7.1 Considerations for Inclusion of Various Stakeholder Groups: Prerequisites for Understanding Tricking Algorithms

Adversarial examples that intentionally attack algorithms by providing malicious input emerge as a notable risk to current black box models [33]. Adversarial examples are commonly seen in the nefarious intent of hackers, leading to security issues [35]. However, this study has shown that laypeople are also deliberately gaming algorithms. Users circumvent translation algorithms by implementing morphological, phonological, optical, and semantic tricks. It should be noted that while hosts cannot directly manipulate or delete the reviews [1], we found that hosts are indirectly managing their reputations by flagging negative reviews and requesting platforms to delete them, or by threatening the reviewers. Our work builds on an increasing number of strategic actions of gaming algorithms [7, 10, 27, 61].

Why do users trick algorithms and how are the stakeholders intertwined around this gaming behavior? We found that while it seems that gaming occurs as a dyadic interaction between a user and the algorithm at a surface level, this game premises an implicit social context where diverse stakeholders interact [10]. Figure 3 illustrates how diverse stakeholder interact around the algorithm. Users circumvent algorithms (reviewer-algorithm) in order to avoid censorship of the host and to mitigate an interpersonal burden (reviewer-host). Furthermore, users trick algorithms to assist other users (reviewer-potential customer) with good intention of building trustworthy online review communities (reviewer-community). In response, users who read encrypted reviews, generally perceive those reviews as more trustworthy and authentic when the negative information is included (potential customer-reviewer).

Our results imply that we should consider the interactions *around* the model to understand human-algorithm interaction [27]. This interaction pattern can be driven by the structural characteristics of the platform that affects content creation, consumption, and moderation [36]. Unlike online communities where there are no restrictions in place for whom cab generate and post the content, a number of peer-to-peer platforms including Airbnb and Amazon allow only authenticated users who are actual, active participants in the related transaction to upload their reviews. In respect of online communities which allow everyone to produce content, companies often employ professional human moderators or use filtering algorithms to filter harmful content or to detect fraudulent content [74]. On the other hand, in the peer-to-peer community in which only the stakeholders who are directly involved in the transaction can write reviews, those stakeholders tend to participate more actively in the content creation and circulation by interacting with and around algorithms as well as platforms. Our findings suggest that, beyond previous findings that hosts actively engage with algorithms [40, 59], reviewers also interact with them and various stakeholders should be considered within this interaction.

Furthermore, the focus of the moderation for the peer-to-peer communities is to encourage users to write and share authentic content that would be helpful for others in reaching a purchasing decision [53]. Indeed, online reviews actually affect people's offline purchasing behavior [57]. Companies are adopting solutions including mutual evaluation systems and translation functions to achieve this goal. However, our study has revealed that these communities do not support an environment in which the users can write authentic reviews. The mass production of such fake reviews occurs not only on platforms that fail to scrutinize reviewers' qualifications such as Google Places or Yelp [57], but also in communities like Airbnb where only real customers can write reviews. A number of users write encrypted reviews to circumvent translation algorithms and censoring of the hosts, leading to information asymmetry. Although we focused on the users' gaming behavior in the Airbnb community, similar phenomena may occur in other global online communities. For example, in terms of online e-commerce websites, users could write reviews in a machine non-translatable way so as to provide authentic information regarding products with deceptive and exaggerated information. However, the interaction pattern in this case could differ from that of Airbnb since the characteristics of the commodities and stakeholders involved in the transaction are dissimilar. Therefore, future work should identify the inherent operation mechanism underpinning the gaming behavior through analysis of various cases across different platforms. Furthermore, this implies that the current global peer-to-peer platforms should adapt a more user-friendly design, which will be discussed in the next section.

7.2 Towards Environment Promoting Reviews for Good: Anonymous and Granular Review Systems

The results of the user study show that reviewers of online booking services tend to trick algorithms behind user interfaces. Airbnb users do not only want to read given reviews as they are, but also to recognize reviewers' hidden intentions behind them. Consequently, encrypted reviews give users more credibility and a sense of belonging. These results call for a discussion of revised design considerations for the online review community and interface. Creating of workarounds and misusing features that interface provides suggests that there is room for improvement in the design of the service [49]. We can think of redesign considerations in stages, focusing first on cause and then solution.

First, it is necessary to present these types of reviews to the users in a more refined and targeted manner. Currently, since AI algorithms cannot correctly distinguish encrypted reviews from the non-encrypted, they are likely to display the reviews without very much distinction or understanding of their intention behind them. Typically, only verified and specialized reviews can be identified separately and placed at the top of a list. Although some users may recognize a reviewer's intention and click the "Like" button to make the review more visible, it is difficult for these reviews to surface from the start without such feedback. To overcome the problem of only these easy-to-translate reviews being placed at the top of the list, of course, improvements of algorithms should be preceded. The various typologies we have discovered and summarized through Study 2 could be utilized for this improvement. In addition, the system can also identify user groups that the review targets, so that the relevant texts are exposed more frequently to them (e.g., in the way to preferentially showing Korean reviews to Korean users thereof).

In addition, ultimately the interface should be designed to create a community where users can leave candid reviews without having to cheat the algorithm. According to our findings and various prior studies, most users of review communities tend to praise hosts or accommodations more than necessary in consideration of their relationships with the hosts or the concern of being evaluated as a guest by the hosts. This reciprocal reviewing practice may hamper them from leaving an objective review, and furthermore will not cultivate a review system that is honest and reliable [58]. We believe that these concerns should be alleviate and propose two approaches as design considerations: anonymous evaluation and multi-dimensional ratings.

First, we can consider introducing an anonymous evaluation to ensure that reviewers do not reveal their identity. This will allow them to become more proactive and to leave candid reviews without worrying about their relationships with the hosts and their reputations. Anonymity enables users to generate more honest and critical reviews since it prevents censorship and protest from the business owner [41, 56]. Of course, potential side effects that may arise from anonymous evaluation should also be addressed. Anonymous evaluation can affect the credibility and perception of product quality [39]. Anonymity can lead individuals to exert less physical effort than those working in an identifiable way [42]. Rather than a fully anonymous evaluation in which reviewers are not identifiable by all community users, it may be sufficient that they are not identifiable only by the hosts. Introducing voting or feedback on reviews by other users may be considered as a viable complement to this consideration. We often see that users' feedback on comments of YouTube video clips motivates the community to respond with more creative and productive comments.

The second consideration is to introduce more fine-grained and multi-dimensional ratings systems that allow users to quantitatively review accommodations and various elements of the host, in addition to the written reviews [8, 12]. Currently, most reviews are rated by a star point system that summarizes an overall facility, and most accommodations tend to receive high scores. Users need to be required to evaluate each of the housing sub-attributes, and this data will provide other users a more objective assessment of the accommodation. These attributes should not be static or fixed, but may be chosen by hosts or reviewers so that they can effectively deliver authentic characteristics of housings.

Third, a system that filters and curates the reviews based on the similar users can engender a better user experience. People trust information from personal sources provided by other human users more than information from impersonal sources [48]. We found that users form bonds with similar linguistic and cultural backgrounds that function as a basis of a trustworthy community.

Furthermore, if the system preferentially curates the reviews written by users who share similar characteristics (i.e., commodity preference, purchase history), users can effectively consume the reviews.

7.3 Integrating Human Intelligence in Algorithms: Considerations for Improving Translation and Moderation Algorithms

Our work can be extended to translation and online moderation since the strategies of deceiving machine translation algorithms is in close concurrence with those of circumventing content moderation and censoring algorithms. Although improving machine translation for the encrypted comments is challenging from a technical point of view as the encryptions are idiosyncratic and contextual, our typology from Study 2 can contribute to the field of online content moderation by providing human's social and linguistic heuristics of evading natural language processing algorithms. Recent advances in natural language processing and sentiment analysis support effective online content moderation at scale [64, 75]. We believe that the human-based heuristics and user-generated examples can help to develop more robust translation and moderation algorithms. Integrating human knowledge or mental model in the machine learning loop can positively influence machine intelligence [32, 34].

In view of data annotation, we can consider the crowdsourcing scheme to label the encrypted reviews generated from peer-to-peer community, which can be used as data sets to build more robust models. Although content such as user-generated review data (e.g., negative information of the assets) and harmful content (e.g., hate speech or harassment) are different, the linguistic and social heuristic of subverting translation and moderation algorithms based on natural language processing can be elaborate. In contrast to labeling hate speech or harassment that are not explicitly aggressive but mentally irritating, labeling those mentally harmless reviews can be a more enjoyable task. A number of participants who read the machine non-translatable review mentioned that "It's quirky and fun," and "I burst into laughter as soon as I read it. Very cheerful reviews."

7.4 Limitations and Future Work

Our work has several limitations. First, given that our participants in Study 3 were all metropolitan Koreans and the data set used in the Study was Korean reviews, the result of this study may not be generalizable. However, it should also be noted that this study does not intend to uncover the full spectrum of interaction patterns with algorithmic systems. Instead, we aim to build on previous literature with this specific empirical case. The behavioral patterns of users avoiding algorithms by modifying language is a global phenomenon [7, 31, 71, 76, 80]. As can be seen from Table 1, the tricking strategies classified in this study can also be applied to other languages. Nevertheless, there are morphological and structural differences respective to each language. In order to categorize strategies precisely according to linguistic characteristics, future work must be conducted utilizing different language samples. Secondly, our study was centered around Airbnb. In future projects, we aim to investigate users' gaming behaviors in other global peer-to-peer platforms such as Amazon, Alibaba, Yelp, and Booking.com. Thirdly, we have not observed hosts' or business owners' perceptions of the encrypted messages. Nevertheless, we tried to infer their perceptions through analyzing the comments in the reviews as described in Section 5.4. We plan to improve our research to cover more diverse stakeholders including hosts, company, and curation algorithms.

8 CONCLUSION

People interact with algorithms in various ways. They not only consume content that algorithms recommend but also actively engage in manipulating algorithms to achieve favorable outcomes. They even game and trick algorithms to interfere with their normal functionalities. This can affect the user experience of various stakeholders involved. We tried to understand how diverse stakeholders interact with each other in tricking algorithms, through a case study of online review communities. Applying a mixed-method approach, we investigated how and why users write machine non-translatable reviews and how those encrypted messages are perceived by the recipients. We learned that users trick the algorithms to avoid censoring, mitigate interpersonal burden, protect privacy, and build informative review communities. We also identified several linguistic and social strategies of writing those reviews. Users perceive encrypted messages as both more trustworthy and authentic. Based on these findings, we discussed implications on the online review community and content moderation algorithms. We hope our work will help the HCI community advance their studies on how humans interact with algorithms.

ACKNOWLEDGMENTS

To Robert, for the bagels and explaining CMYK and color spaces.

REFERENCES

[1] Airbnb. 2020. Can I remove or respond to a review I disagree with? - Airbnb Help Center. https://www.airbnb.com/help/article/32/can-i-remove-or-respond-to-a-review-i-disagree-with?_set_bev_on_new_domain=1609918482_OTYxODRmODUwMGNm

[2] Nisamar Baute-Díaz, Desiderio Gutiérrez-Taño, and Ricardo J Díaz-Armas. 2020. What drives guests to misreport their experiences on Airbnb? A structural equation modelling approach. Current Issues in Tourism (2020), 1–18.

[3] Sophie Bishop. 2019. Managing visibility on YouTube through algorithmic gossip. New media & society 21, 11-12 (2019), 2589–2606.

[4] Katherine Blashki and Sophie Nichol. 2005. Game geek’s goss: linguistic creativity in young males within an online university forum (94/\/\3 933k’5 9055oneone). Australian journal of emerging technologies and society 3, 2 (2005), 71–80.

[5] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. Qualitative research in psychology 3, 2 (2006), 77–101.

[6] Judith Bridges and Camilla Vásquez. 2018. If nearly all Airbnb reviews are positive, does that make them meaningless? Current Issues in Tourism 21, 18 (2018), 2057–2075.

[7] Jenna Burrell, Zoe Kahn, Anne Jonas, and Daniel Griffin. 2019. When Users Control the Algorithms: Values Expressed in Practices on Twitter. Proceedings of the ACM on Human-Computer Interaction 3, CSCW (2019), 1–20.

[8] Pei-Yu Chen, Yili Hong, and Ying Liu. 2018. The value of multidimensional rating systems: Evidence from a natural experiment and randomized experiments. Management Science 64, 10 (2018), 4629–4647.

[9] Jeung-Ryeul Cho and Hsuan-Chih Chen. 1999. Orthographic and phonological activation in the semantic processing of Korean Hanja and Hangul. Language and Cognitive Processes 14, 5-6 (1999), 481–502.

[10] Kelley Cotter. 2019. Playing the visibility game: How digital influencers and algorithms negotiate influence on Instagram. New Media & Society 21, 4 (2019), 895–913.

[11] Benjamin Edelman. 2017. The market design and policy of online review platforms. Oxford Review of Economic Policy 33, 4 (2017), 635–649.

[12] Malin Eiband, Sarah Theres Völkel, Daniel Buschek, Sophia Cook, and Heinrich Hussmann. 2019. When people and algorithms meet: User-reported problems in intelligent everyday applications. In Proceedings of the 24th International Conference on Intelligent User Interfaces. 96–106.

[13] Ziv Epstein, Gordon Pennycook, and David Rand. 2020. Will the crowd game the algorithm? Using layperson judgments to combat misinformation on social media by downranking distrusted sources. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. 1–11.

[14] Motahhare Eslami, Karrie Karahalios, Christian Sandvig, Kristen Vaccaro, Aimee Rickman, Kevin Hamilton, and Alex Kirlik. 2016. First I" like" it, then I hide it: Folk Theories of Social Feeds. In Proceedings of the 2016 cHI conference on human factors in computing systems. 2371–2382.

[15] Motahhare Eslami, Kristen Vaccaro, Karrie Karahalios, and Kevin Hamilton. 2017. "Be Careful; Things Can Be Worse than They Appear": Understanding Biased Algorithms and Users’ Behavior Around Them in Rating Platforms.. In ICWSM. 62–71.

[16] Motahhare Eslami, Kristen Vaccaro, Karrie Karahalios, and Kevin Hamilton. 2017. “Be careful; things can be worse than they appear”: Understanding Biased Algorithms and Users’ Behavior around Them in Rating Platforms. In Proceedings of the International AAAI Conference on Web and Social Media, Vol. 11.

[17] Motahhare Eslami, Kristen Vaccaro, Min Kyung Lee, Amit Elazari Bar On, Eric Gilbert, and Karrie Karahalios. 2019. User attitudes towards algorithmic opacity and transparency in online reviewing platforms. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. 1–14.

[18] Raffaele Filieri. 2016. What makes an online consumer review trustworthy? Annals of Tourism Research 58 (2016), 46–64.

[19] Kristopher Floyd, Ryan Freling, Saad Alhoqail, Hyun Young Cho, and Traci Freling. 2014. How online product reviews affect retail sales: A meta-analysis. Journal of Retailing 90, 2 (2014), 217–232.

[20] Kenneth I Forster, Chris Davis, Colin Schoknecht, and Ronald Carter. 1987. Masked priming with graphemically related forms: Repetition or partial activation? The Quarterly Journal of Experimental Psychology 39, 2 (1987), 211–251.

[21] Ge Gao, Hao-Chuan Wang, Dan Cosley, and Susan R Fussell. 2013. Same translation but different experience: the effects of highlighting on machine-translated conversations. In Proceedings of the sigchi conference on human factors in computing systems. 449–458.

[22] Ge Gao, Bin Xu, Dan Cosley, and Susan R Fussell. 2014. How beliefs about the presence of machine translation impact multilingual collaborations. In Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing. 1549–1560.

[23] Sanjam Garg, Craig Gentry, Amit Sahai, and Brent Waters. 2013. Witness encryption and its applications. In Proceedings of the forty-fifth annual ACM symposium on Theory of computing. 467–476.

[24] Oliver Gee. 2019. Should foreigners steer clear of France's 'backwards language' Verlan? https://www.thelocal.fr/20190510/should-foreigners-steer-clear-of-frances-backwards-language

[25] Tarleton Gillespie. 2017. Algorithmically recognizable: Santorum's Google problem, and Google's Santorum problem. Information, communication & society 20, 1 (2017), 63–80.

[26] Christine Guerrera and Kenneth Forster. 2008. Masked form priming with extreme transposition. Language and Cognitive Processes 23, 1 (2008), 117–142.

[27] Jesse Haapoja, Salla-Maaria Laaksonen, and Airi Lampinen. 2020. Gaming Algorithmic Hate-Speech Detection: Stakes, Parties, and Moves. Social Media+ Society 6, 2 (2020), 2056305120924778.

[28] Jack Halpern. 2002. Lexicon-based orthographic disambiguation in CJK intelligent information retrieval. In COLING-02: The 3rd Workshop on Asian Language Resources and International Standardization.

[29] Jeffrey T Hancock, Mor Naaman, and Karen Levy. 2020. AI-Mediated Communication: Definition, Research Agenda, and Ethical Considerations. Journal of Computer-Mediated Communication 25, 1 (2020), 89–100.

[30] Eric Donald Hirsch, Joseph F Kett, and James S Trefil. 2002. The new dictionary of cultural literacy. Houghton Mifflin Harcourt.

[31] Chaya Hiruncharoenvate, Zhiyuan Lin, and Eric Gilbert. 2015. Algorithmically Bypassing Censorship on Sina Weibo with Nondeterministic Homophone Substitutions.. In ICWSM. Citeseer, 150–158.

[32] Andreas Holzinger, Markus Plass, Michael Kickmeier-Rust, Katharina Holzinger, Gloria Cerasela Crişan, Camelia-M Pintea, and Vasile Palade. 2019. Interactive machine learning: experimental evidence for the human in the algorithmic loop. Applied Intelligence 49, 7 (2019), 2401–2414.

[33] Ling Huang, Anthony D Joseph, Blaine Nelson, Benjamin IP Rubinstein, and J Doug Tygar. 2011. Adversarial machine learning. In Proceedings of the 4th ACM workshop on Security and artificial intelligence. 43–58.

[34] Ting-Hao Huang, Joseph Chee Chang, and Jeffrey P Bigham. 2018. Evorus: A crowd-powered conversational assistant built to automate itself over time. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. 1–13.

[35] Matthew Hutson. 2018. Hackers easily fool artificial intelligences.

[36] High Tech Law Institute. 2018. Conference of Content Moderation & Removal at Scale. https://law.scu.edu/event/content-moderation-removal-at-scale/

[37] iProperty Management & Investments. 2020. Airbnb Statistics 2020: User & Market Growth Data. https://ipropertymanagement.com/research/airbnb-statistics

[38] Maurice Jakesch, Megan French, Xiao Ma, Jeffrey T Hancock, and Mor Naaman. 2019. AI-mediated communication: How the perception that profile text was written by AI affects trustworthiness. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. 1–13.

[39] Matthew L Jensen, Joshua M Averbeck, Zhu Zhang, and Kevin B Wright. 2013. Credibility of anonymous online product reviews: A language expectancy perspective. Journal of Management Information Systems 30, 1 (2013), 293–324.

[40] Shagun Jhaver, Yoni Karpfen, and Judd Antin. 2018. Algorithmic anxiety and coping strategies of Airbnb hosts. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. 1–12.

[41] Ruogu Kang, Stephanie Brown, and Sara Kiesler. 2013. Why do people seek anonymity on the internet? Informing policy and design. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 2657–2666.

[42] Norbert L Kerr and Steven E Bruun. 1981. Ringelmann revisited: Alternative explanations for the social loafing effect. Personality and social psychology bulletin 7, 2 (1981), 224–231.

[43] Soomin Kim, Jinsu Eun, Changhoon Oh, Bongwon Suh, and Joonhwan Lee. 2020. Bot in the Bunch: Facilitating Group Chat Discussion by Improving Efficiency and Participation with a Chatbot. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. 1–13.

[44] Taewook Kim, Jung Soo Lee, Zhenhui Peng, and Xiaojuan Ma. 2019. Love in Lyrics: An Exploration of Supporting Textual Manifestation of Affection in Social Messaging. Proceedings of the ACM on Human-Computer Interaction 3, CSCW (2019), 1–27.

[45] Eliscia Kinder, Mohammad Hossein Jarrahi, and Will Sutherland. 2019. Gig Platforms, Tensions, Alliances and Ecosystems: An Actor-Network Perspective. Proceedings of the ACM on Human-Computer Interaction 3, CSCW (2019), 1–26.

[46] Robert Allen King, Pradeep Racherla, and Victoria D Bush. 2014. What we know and don't know about online word-of-mouth: A review and synthesis of the literature. Journal of interactive marketing 28, 3 (2014), 167–183.

[47] Neha Kumar and Nimmi Rangaswamy. 2013. The mobile media actor-network in urban India. In Proceedings of the SIGCHI conference on human factors in computing systems. 1989–1998.

[48] Johannes Kunkel, Tim Donkers, Lisa Michael, Catalin-Mihai Barbu, and Jürgen Ziegler. 2019. Let me explain: impact of personal and impersonal explanations on trust in recommender systems. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. 1–12.

[49] Airi Lampinen and Barry Brown. 2017. Market design for HCI: Successes and failures of peer-to-peer exchange platforms. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. 4331–4343.

[50] Bruno Latour. 1996. On actor-network theory: A few clarifications. Soziale welt (1996), 369–381.

[51] Eun-Ju Lee, Hye-Yon Lee, and Sukyoung Choi. 2020. Is the message the medium? How politicians' Twitter blunders affect perceived authenticity of Twitter communication. Computers in Human Behavior 104 (2020), 106188.

[52] Natalie J Lefkowitz. 1989. Verlan: talking backwards in French. The French Review 63, 2 (1989), 312–322.

[53] Hanlin Li and Brent Hecht. 2021. 3 Stars on Yelp, 4 Stars on Google Maps: A Cross-Platform Examination of Restaurant Ratings. Proceedings of the ACM on Human-Computer Interaction 4, CSCW3 (2021), 1–25.

[54] Daniel J Liebling, Michal Lahav, Abigail Evans, Aaron Donsbach, Jess Holbrook, Boris Smus, and Lindsey Boran. 2020. Unmet Needs and Opportunities for Mobile Translation AI. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. 1–13.

[55] Hajin Lim, Dan Cosley, and Susan R Fussell. 2018. Beyond Translation: Design and Evaluation of an Emotional and Contextual Knowledge Interface for Foreign Language Social Media Posts. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. 1–12.

[56] Ruiling Lu and Linda Bol. 2007. A comparison of anonymous versus identifiable e-peer review on college student writing performance and the extent of critical feedback. Journal of Interactive Online Learning 6, 2 (2007).

[57] Michael Luca. 2016. Reviews, reputation, and revenue: The case of Yelp. com. Com (March 15, 2016). Harvard Business School NOM Unit Working Paper 12-016 (2016).

[58] Michael Luca. 2017. Designing online marketplaces: Trust and reputation mechanisms. Innovation Policy and the Economy 17, 1 (2017), 77–93.

[59] Michael Luca and Georgios Zervas. 2016. Fake it till you make it: Reputation, competition, and Yelp review fraud. Management Science 62, 12 (2016), 3412–3427.

[60] Xiao Ma, Justin Cheng, Shankar Iyer, and Mor Naaman. 2019. When Do People Trust Their Social Groups?. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. 1–12.

[61] Xiao Ma, Jeffery T Hancock, Kenneth Lim Mingjie, and Mor Naaman. 2017. Self-disclosure and perceived trustworthiness of Airbnb host profiles. In Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing. 2397–2409.

[62] Jo Mackiewicz and Dave Yeats. 2014. Product review users' perceptions of review quality: The role of credibility, informativeness, and readability. IEEE Transactions on Professional Communication 57, 4 (2014), 309–324.

[63] Raveesh Mayya, Shun Ye, Siva Viswanathan, and Rajshree Agarwal. 2019. Who Forgoes Screening in Online Markets and When? Evidence from Airbnb. Evidence from Airbnb (March 2019) (2019).

[64] Jihyung Moon, Won Ik Cho, and Junbum Lee. 2020. BEEP! Korean Corpus of Online News Comments for Toxic Speech Detection. In Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media. Association for Computational Linguistics, Online, 25–31. https://www.aclweb.org/anthology/2020.socialnlp-1.4

[65] Alfred Ng. 2020. Teens have figured out how to mess with Instagram's tracking algorithm. https://www.cnet.com/news/teens-have-figured-out-how-to-mess-with-instagrams-tracking-algorithm/

[66] Varvara Nikulina, Johan Larson Lindal, Henrikke Baumann, David Simon, and Henrik Ny. 2019. Lost in translation: A framework for analysing complexity of co-production settings in relation to epistemic communities, linguistic diversities and culture. Futures 113 (2019), 102442.

[67] Changhoon Oh, Taeyoung Lee, Yoojung Kim, SoHyun Park, Saebom Kwon, and Bongwon Suh. 2017. Us vs. them: Understanding artificial intelligence technophobia over the google deepmind challenge match. In Proceedings of the

2017 CHI Conference on Human Factors in Computing Systems. 2523–2534.

[68] Lee-Yun Pan and Jyh-Shen Chiou. 2011. How much can you trust online information? Cues for perceived trustworthiness of consumer-generated online information. Journal of Interactive Marketing 25, 2 (2011), 67–74.

[69] Mei-Hua Pan, Naomi Yamashita, and Hao-Chuan Wang. 2017. Task Rebalancing: Improving Multilingual Communication with Native Speakers-Generated Highlights on Automated Transcripts. In Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing. 310–321.

[70] Manuel Perea and Stephen J Lupker. 2004. Can CANISO activate CASINO? Transposed-letter similarity effects with nonadjacent letter positions. Journal of memory and language 51, 2 (2004), 231–246.

[71] Sherisse Pham. 2018. Chinese censors are scanning WeChat images to block banned words. https://money.cnn.com/2018/03/01/technology/china-wechat-censorship-ai/index.html

[72] Rob Price. 2019. People are slipping fake baby and marriage announcements into Facebook posts to trick the algorithm into boosting their posts. https://businessinsider.com/facebook-users-claim-pregnant-married-trick-algorithm-boost-posts-2019-11

[73] Emilee Rader and Rebecca Gray. 2015. Understanding user beliefs about algorithmic curation in the Facebook news feed. In Proceedings of the 33rd annual ACM conference on human factors in computing systems. 173–182.

[74] Sarah T Roberts. 2019. Behind the screen: Content moderation in the shadows of social media. Yale University Press.

[75] Keisuke Sakaguchi, Kevin Duh, Matt Post, and Benjamin Van Durme. 2016. Robsut wrod reocginiton via semi-character recurrent neural network. arXiv preprint arXiv:1608.02214 (2016).

[76] Jeanette Si. 2018. The Chinese Language as a Weapon: How China's Netizens Fight Censorship. https://medium.com/berkman-klein-center/the-chinese-language-as-a-weapon-how-chinas-netizens-fight-censorship-8389516ed1a6

[77] Bogusia Temple and Alys Young. 2004. Qualitative research and translation dilemmas. Qualitative research 4, 2 (2004), 161–178.

[78] José Van Dijck. 2013. The culture of connectivity: A critical history of social media. Oxford University Press.

[79] Sarah Theres Völkel, Renate Haeuslschmid, Anna Werner, Heinrich Hussmann, and Andreas Butz. 2020. How to Trick AI: Users' Strategies for Protecting Themselves from Automatic Personality Assessment. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. 1–15.

[80] Audrey M Wozniak. 2015. River-crabbed Shitizens and missing knives: A sociolinguistic analysis of trends in Chinese language use online as a result of censorship. Applied Linguistics Review 6, 1 (2015), 97–120.

[81] Kyung Hyan Yoo and Ulrike Gretzel. 2008. What motivates consumers to write online travel reviews? Information Technology & Tourism 10, 4 (2008), 283–295.

[82] Malte Ziewitz. 2019. Rethinking gaming: The ethical work of optimization in web search engines. Social studies of science 49, 5 (2019), 707–731.

Received January 2021; revised April 2021; accepted May 2021