Wang, Soyoung

Conference Paper

How will users respond to the adversarial noise that prevents the generation of deepfakes?

23rd Biennial Conference of the International Telecommunications Society (ITS): "Digital societies and industrial transformations: Policies, markets, and technologies in a post-Covid world", Online Conference / Gothenburg, Sweden, 21st-23rd June, 2021

Provided in Cooperation with:

International Telecommunications Society (ITS)

Suggested Citation: Wang, Soyoung (2021) : How will users respond to the adversarial noise that prevents the generation of deepfakes?, 23rd Biennial Conference of the International Telecommunications Society (ITS): "Digital societies and industrial transformations: Policies, markets, and technologies in a post-Covid world", Online Conference / Gothenburg, Sweden, 21st-23rd June, 2021, International Telecommunications Society (ITS), Calgary

This Version is available at:
https://hdl.handle.net/10419/238060

How will users respond to the adversarial noise that prevents the generation of deepfakes?[1]

Soyoung Wang*

* Korea University, South Korea, soyoung1007@korea.ac.kr

Abstract

The development of artificial intelligence (AI) technology has made it easy for users to generate hyper-realistic fake media content, and its most representative by-product is called deepfake. However, considerable attention has been paid to the adverse effects of deepfakes as they are tightly connected to the production of fake news, financial frauds, or fake pornographies. The misuse of deepfakes led to a series of deepfake prevention studies, but most were post-detection methods. This study thus investigated deepfake malfunction-inducing technology that may forestall the generation of deepfake with PGD attack. In the next part of the study, overall preferences and intention to use were measured as people's responses to this technology. An online survey especially targeting those exposed to various media like social media influencers (SMIs), was conducted. The deepfakes started to malfunction after adding 0.009 levels of an adversarial noise as a preventive mechanism. From a technical viewpoint, higher noise was a more effective way to prevent deepfake synthesis, but from the user's viewpoint, noise as high as 0.03 was found to be appropriate. Individuals' intention to use was tested with Bulgurcu's ISP compliance model. It was found that SMIs' predictive evaluations on the cost and benefit of this technology influence their attitude, and consequently, their intention to use it. This study shows the value of collaborative studies of AI-based privacy security domain and media industry domain. It also expands the scope of the framework with thorough hypothetical testing in the deepfake context.

Keywords: Deepfake, Adversarial noise, Image quality, Intention to use

1. Introduction

The advancement of emerging artificial intelligence (AI) technology has led to the rapid development of image synthesis, editing, or style transformation (Chesney & Citron, 2019). Deepfakes are one of the products of advanced AI algorithms that can transpose one's face image onto an existing photo or video (Bates, 2018; Westerlund, 2019). Although some benefits of the deepfakes, considerable attention is being paid to the negative consequences of deepfakes nowadays because they are deeply related to the production of fake news, financial fraud threats, or fake pornography of celebrities. In addition, deepfakes pose risks to those being deep faked and those being misled to believe deepfake.

Such misuse of deepfakes has led to various research on deepfake prevention, but inherently reactive and mostly post-detection methods (Bappy et al., 2019; Korshunov & Marcel, 2018; Nguyen

Acknowledgment This work was supported by the Ministry of Education of the Republic of Korea, the National Research Foundation of Korea (NRF-2019S1A3A2099973), and the MSIT (Ministry of Science and ICT) Korea, under the ITRC (Information Technology Research Center) support program (IITP-2020-0-01749) supervised by the IITP (Institute of Information & Communications Technology Planning & Evaluation).

et al., 2019; Westerlund, 2019). Given that almost 96 percent of deepfake media are illegal pornography of female celebrities (Ajder et al., 2019), research into baseline measures as preemptive prevention of deepfake synthesis is also necessary and valuable. Further, recent studies on deepfakes have mainly concentrated on explaining the concept and relevant topic of the deepfake technology and products (Bae, 2019; Delfino, 2019; Fletcher, 2018; Gosse & Burkell, 2020; Harper et al., 2021; Kietzmann et al., 2019) or the legal issues behind the deepfake phenomenon (Giseke, 2020; Hall, 2018; Harris, 2018; Maras & Alexandrou, 2019). In contrast, there is a lack of research on users who lead an online ecosystem and can either become creators, viewers, or victims of the deepfake. Moreover, research on users' responses to the obviate mechanisms to the deepfake can provide valuable knowledge for identifying individuals' likely attitudes toward new information technology (IT), which are significant for explaining successful acceptance or resistance of emerging IT (Park et al., 2016).

In this respect, this study has two objectives primarily. One is to explore the forestalled model to prevent deepfake using projected gradient descent (PGD). This adversarial attack algorithm can find an adversarial noise that can malfunction deepfake neural networks. The other is to investigate people's satisfaction and intention to use this adversarial noise as the preemptive defense method. Image quality criteria were borrowed from photographic research. In addition, Bulgurcu's ISP compliance model was adopted to explore overall determinants that affect an individual's attitude toward the noise and subsequent usage intention. Consequently, this study attempts to present the usefulness of a wide range of multidisciplinary studies in a hyper-convergence society and the necessity for user study based on a solid framework.

2. Literature review and research hypotheses

2.1. Deepfake technology

The development of generative models like *generative adversarial network (GAN)*, made by artificial neural *network* or deep learning, has advanced the performance of image creation, image transformation, or style conversion. As can be inferred from the abbreviation, one of the networks within the GAN *generates* real-like fake data that is indistinguishable from the original one, called a generator. On the other hand, receiving the hyper-realistic data from the generator, a discriminator tries to distinguish them from the authentic to the manipulated (Creswell et al., 2018). This simultaneous and competitive process is referred to *as adversarial* training, which is the core idea of GAN and ultimately leads to image manipulation (Westerlund, 2019). Using such a generative model like GAN, algorithms that naturally recombine personal faces and features of a character by image to image translation or face swap are called deepfake.

2.2. Assessing the optimized levels of adversarial noise

2.2.1. At a technical level

Although deepfake based genuine-like image generation is highly accurate, the adversarial attack has shown that artificial neural networks are surprisingly vulnerable to just a minimal level of adversarial noise (hereafter 'noise') to an original image. This noise makes the AI classifier misclassify a subject, leading to a failure of the image synthesis (Szegedy et al., 2013). Namely, the adversarial attack is used to interfere with deepfake algorithms through the fast gradient sign method (FGSM) or projected gradient descent (PGD). These typical adversarial attack algorithms prevent deepfake production by attacking the vulnerability of deepfake algorithms (Madry et al., 2017). Notably, the main goal of the PGD is to generate an adversarial example, which is a noise-added image that appears to be just ordinary in human eyes but has a protective function to stop the image synthesis. However, as the noise is added to the image, adversarial examples are followed by inevitable image quality degradation.

According to the experiment conducted with fellow researchers[2], when zero noise means the original image that can still be manipulated by deepfake, 0.001 level of noise was not enough to significantly adequate for deepfake prevention. However, deepfake has been noticeably hindered from the noise 0.009. Moreover, when the noise level was above 0.1, the result of the deepfake was visually unrecognizable. These experiments have identified that the higher the noise level, the greater the likelihood of preventing image synthesis and transformation by deepfake. In contrast, this study has also confirmed that excessive noise levels, such as 0.5, can severely damage the original image.

Given that PGD will be unnecessary in practice if the image quality downgrades too severely, conducting additional experiments to identify how high the noise level is apt for the effective declination of deepfake generation while maintaining the image quality appears profoundly essential. Accordingly, examining the optimized levels of the noise that does not disturb the identification of the original media is the key to deepfake production prevention.

[2] This part of the study was co-worked with Korea University School of cybersecurity

Image 1: Deepfake malfunction induction experiment results based on the noise level

2.2.2. At a potential user level

The image with the adversarial noise also has to evaluate its' quality with the interplay of diverse attributes. The potential users may form an integrated impression and preference of the noise-added image by comparing it with the original image. The image's tones, resolution, balance, or noise that appear to cross the borderline of subjective and objective measures can be the principal evaluation criteria for the image quality, as previous scholars indicated (Keelan, 2002; Pedersen et al., 2009; Noh & Har, 2010; Radun et al., 2008; Yendrikhovskij et al., 1998). As such, this study suggests that the tones of the picture with subjective evaluation criteria of clearness(cloudiness), cleanness(dirtiness), and brightness(darkness) and other constructs such as resolution, balance, noise, and preference of the picture are the key factors to be evaluated from a user perspective. In this study, the aggregation of measures is predicted to show the individual's overall satisfaction with the image.

Table1: Image quality assessment

		< ---------- >							
		1	**2**	**3**	**4**	**5**	**6**	**7**	
Tones	Cloudedness								Clearness
	Dirtiness								Cleanness
	Darkness								Brightness
Resolution	Unsharpness								Sharpness
Balance	Unnaturalness								Naturalness
Noise	Roughness								Softness
Preference	Bad								Good

The noise levels this study mainly evaluated to see the user response are 0.009 level, 0.03 level, and 0.5 level. In an algorithm aspect, this study discovered that the first noticeable obstruction of deepfake synthesis resulted from the value of 0.009. Then, it was from 0.1 that the deepfake were

completely malfunctioned to an unrecognizable extent. Therefore, we would like to start from a measurement of 0.009- level, which is a minimum level of deepfake prevention, 0.03- level as the medium value, and 0.5 as the level that deepfake attack is not at all possible.

2.3. Information security policy (ISP) compliance model

As information security risks are viewed as significant threats for many organizations, they invest to solutions to reduce such risks. However, some studies have emphasized the roles and responsibilities of an employee, who can become both an abuser and a guardian for a company resource (Myyry et al., 2009; Willison, 2006). Discovering their vital roles in enhancing information security, encouraging them to comply with the information security policy (ISP) has become a central business to organizations (Bulgurcu et al., 2010; Myyry et al., 2009; Pahnila et al., 2007). Particularly, Bulgurcu et al. (2010) integrated two significant theories to develop an ISP compliance model that illustrates antecedents of an employee's compliance with the ISP and the organization. First, drawing on the theory of planned behavior (TPB), the authors contended that an individual's *attitude* toward ISP compliance, as well as *normative beliefs* and *self-efficacy*, provoke an *intention to comply with the ISP*. Next, Bulgurcu et al. (2010) added the rational choice theory (RCT) to investigate factors that may influence an individual's attitude toward ISP compliance. These scholars proposed that an employee's beliefs on whether or not performing the compliance behavior result in specific outcomes (i.e., costs and benefits). These are the determinants of an individual's *attitude* toward compliance behavior (Paternoster & Pogarsky, 2009). The authors defined these beliefs as *beliefs about the overall assessment of compliance and noncompliance*. Moreover, they indicated that an individual's cost/benefit implications about potential outcomes of compliance and noncompliance might come from their perceptions or feelings about their performance. These perceptions are defined as *beliefs about outcomes* in Bulgurcu's ISP compliance model.

Correspondingly, this study adopts Bulgurcu's ISP compliance model as the theoretical framework. Although Bulgurcu's study emphasizes an individual's mandatory or voluntary behavior about an organization's resources, this study delves into an individual's voluntary intention to use the noise as a safeguard to his or her information resources on social media. The study by Taneja et al. (2014) is one example that analyzed an individual's adoption of privacy controls for protecting personal information on Facebook with the framework of the ISP compliance model.

2.4. Research hypotheses

This study suggests that many individuals suffer from serious information security risks with

artificial intelligence (AI) technology development. Deepfake algorithms, for example, can collect photos uploaded by users on social media and synthesize them with other subjects' images or videos. However, there are no proper privacy controls to protect individuals' resources, particularly image or video content, from deepfake attacks so far.

However, the last part of the study has provided the possibility of protecting an individual's resources from deepfake technology through noise. Owing to this potential of noise-based technology, this study examines an individual's likely intention to use noise as one of the safeguard mechanisms in reliance on Bulgurcu's ISP compliance model. In this context, *noise* is defined as image protection controls from deepfake technology used in social media platforms.

2.4.1. Intention to use the noise

The TPB explains an individual's intention toward a given behavior (Fishbein & Ajzen, 1975). Consistent with the extant literature, this study posits that an individual's readily intention to use the noise on social media is associated with attitude, social norm, and perceived behavioral controls. According to Fishbein and Ajzen (1975), attitude is an individual's degree of evaluation of the performance of the behavior, which is positively or negatively valued. Previous scholars have shown that the intention to perform a behavior can be predicted with high accuracy from the *attitude* toward the behavior. *Social norm* refers to the perceived social pressure to perform or not perform in a behavior (Ajzen, 1991). The TPB predicts that if a norm of his or her reference group is friendly, an individual's intention to behavior would also increase. However, the results for social norm are mixed in that some technology acceptance model (TAM) based research have discovered limited findings (Mathieson, 1991; Venkatesh et al., 2003) while others have argued its' explanatory power in individuals' decisions (Christofides et al., 2009; Taneja et al., 2014). *Perceived behavioral control (PBC)* is self-efficacy regarding the behavior (Ajzen, 1991). Prior studies have indicated that she may shy away from using when an individual is not confident in using the technology (Taneja et al., 2014).

Based on previous literature that has examined the constructs of TPB, this study proposes that if an individual has a positive attitude, perceives the norm with clear social expectation, and believes their self-efficacy to control the behavior, he is more likely to have an intention to use the noise available on social media.

H1: Attitude toward the noise is positively related to the intention to use the noise.

H2: Social norms about the noise are positively related to the intention to use the noise.

H3: Perceived behavioral controls toward the noise are positively related to the intention to use the noise.

2.4.2. Beliefs about overall assessment of consequences of using or not using the noise

Amongst the three variables that are perceived to influence the intention to use the noise, an individual's *attitude* toward a behavior is found to be shaped by his or her beliefs about behavior-related results (Ajzen, 1991; Fishbein & Ajzen, 1975). These *beliefs about the overall assessment of consequences* are often accompanied by the cost and benefit judgments of behavior, as Bulgurcu et al. (2010) argued. Similarly, this study posits that the *perceived benefit of using the noise*, *perceived cost of not using the noise*, and *perceived cost of using the noise* are three distinct fundamental beliefs related to an individual's attitude toward using the noise. To be more specific, each belief is the overall prediction of the benefit of using, unexpected cost of not using, and expected cost of using the noise. The expectancy-value theory of attitude suggests that an individual is likely to favor behaviors that lead to desirable outcomes and choose the ones with the most potent combination of expected value (Taneja et al., 2014; Wigfield, 1994). In this study's context, if an individual acknowledges the value of using the noise or the disadvantage of not using the noise, a positive attitude toward using the noise will be formed. Meanwhile, if an individual feels that using the noise is too costly to meet the security requirements, a negative attitude will be revealed. Therefore,

H4: Perceived benefit of using the noise is positively related to attitude toward using the noise.

H5: Perceived cost of not using the noise is positively related to attitude toward using the noise.

H6: Perceived cost of using the noise is negatively related to attitude toward the noise.

2.4.3. Drivers of beliefs about overall assessment of consequences of using or not using the noise

According to Bulgurcu et al. (2010), individuals' overall assessment of the beliefs is influenced by their cognitive processing of beliefs about outcomes. For example, if an employee believes that he will receive a compliment for his ISP complying behavior, it can become the outcome that follows from the compliance. In this context, *beliefs about outcomes* are defined as beliefs that some events will follow from using the noise. Bulgurcu et al. (2010) and Taneja et al. (2014) have examined individuals' salient outcome beliefs that lead to the perceived benefit of, cost of, and cost of not compliance.

Based on this research, this study tries to deploy an ISP compliance model to predict some drivers that can affect an individual's integrated assessments of the benefit and cost of the noise. For example, *intrinsic benefit* and *resource safety* are the two outcome beliefs that are likely to be associated with perceiving the benefit of using the noise. Intrinsic benefits are individuals' preferences and desires that motivate them to perform a behavior (Taylor & Todd, 1995). In the context of using the noise for deepfake prevention, individuals may feel positive feelings if they believe their photos uploaded on social media are well protected from any damage by using the noise. Resource safety is an individual's

perception that his or her information and resources are well-safeguarded by complying with the policy (Bulgurcu et al., 2010). In this study's context, individuals may likely feel safe and perceive the benefit of using the noise when they think the noise can eliminate or at least reduce potential damage or misuse of their images or videos.

H7: Intrinsic benefit is positively related to the perceived benefit of using the noise.

H8: Resource safety is positively related to the perceived benefit of using the noise.

Second, *threat severity, resource vulnerability,* and *privacy risk* are the three outcome beliefs that drive individuals' assessment of the cost of not following the behavior. *Threat severity* is an individual's beliefs on the magnitude of the threat. Privacy-related literature has shown that a severe threat of privacy invasion can arouse a fear appeal that leads to an individual's engagement in protective behavior (Rogers, 1975; Johnston & Warkentin, 2010). In the deepfake context, the more severe individuals recognize the threat of private image abuse, the more the expected cost for not using the noise is likely to be high. *Resource vulnerability* is an individual's belief that his or her resources will be susceptible to harm or threats, according to the PMT (Rogers, 1975). Taneja et al. (2014) has discovered that resource vulnerability contributes to higher recognition of the cost of not using the privacy controls on Facebook. Consistent with the literature, this study postulates that individuals may perceive the higher cost of not using the noise if they think their images or videos as personal resources are vulnerable to potential misuse by unauthorized users. *Privacy risk* is a salient belief in privacy contexts, which is defined as the expectation of losses resulting from disclosing personal information (Li et al., 2010; Xu et al., 2008). As the risk of personal losses decreases users' willingness to post the content about themselves on social media, individuals may use the noise to avoid the risk of illegal exploitation of deepfake.

H9: Threat severity is positively related to the perceived cost of not using the noise.

H10: Resource vulnerability is positively related to the perceived cost of not using the noise.

H11: Privacy risk is positively related to the perceived cost of not using the noise.

Lastly, *work impediment* and *resource distortion* are the two outcome beliefs that drive an individual's evaluation of the cost of acting. *Work impediment* is an individual's belief that performing the behavior will harm his tasks or activities (Bulgurcu et al., 2010). Viane et al. (2004) found that if individuals feel that their goal or outcome is disrupted by performing the course of action, they may feel unpleasant and react to avoid such conduct. Thus, the higher the work impediment predicts the

higher cost of using the noise. In line with this context, *resource distortion* can also be the driver of avoiding the behavior. Resource distortion is an individual's belief that his or her content may be distorted or blurred as a consequence of using the noise. If individuals' goals to maintain content quality are disturbed by adding the noise, they may feel unpleasant. For instance, the noise can be perceived as a cost when the overall mood of the original image, consisting of color, sharpness, or clarity, is distorted from adopting it. Therefore, the higher the resource distortion, the higher the cost of using the noise. The summarized research model is presented in Figure 2.

H12: Work impediment is positively related to the perceived cost of using the noise.

H13: Resource distortion is positively related to the perceived cost of using the noise.

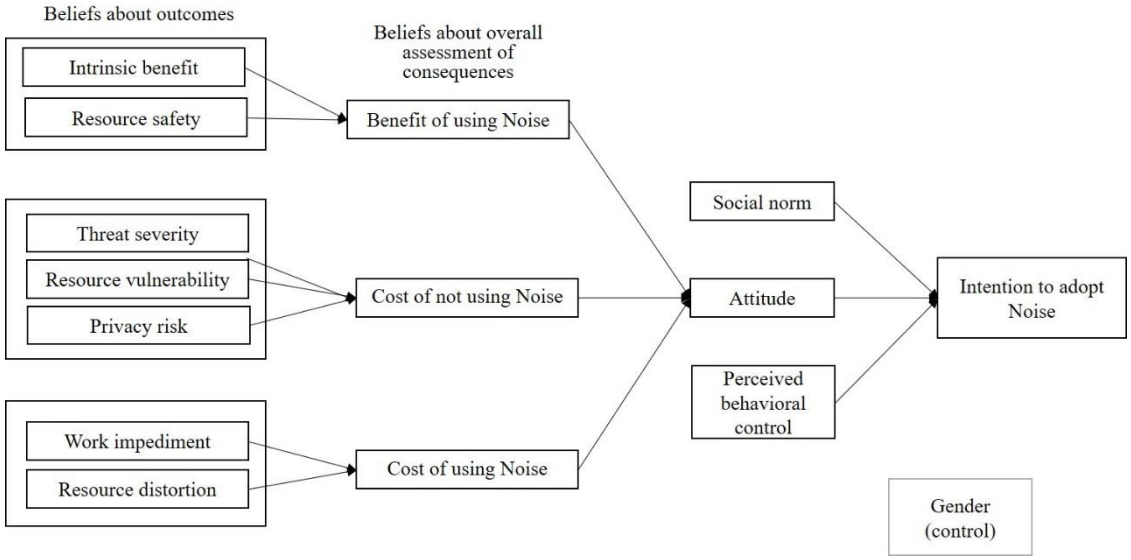

Figure 2. Research Model

3. **Methodology**

3.1. Data collection

This study used an online survey to test the research model. The survey was conducted from May 28 to June 4, 2021, in Korea. This study specifically addressed social media influencers (SMIs) as the main subjects. A snowball sampling was used to collect the data. The sampling criteria included either (1) social media influencers who have more than 1,000 followers or subscribers; or (2) individuals who are with jobs that their faces be constantly exposed to TV, broadcasting, or any online platforms; and (3) those with aged between their 20s up to 30s. This study predicted that individuals whose self-appearance is consistently posted and shared through media would be more aware of the risks of AI

technologies on their private resources. In addition, since the noise is not yet emerged, this study had to find samples that may understand the necessity of privacy controls and predict likely beliefs about using them. Given its' predictive nature, SMIs have been selected as samples for this research.

As such, a total of 96 samples was collected, and with removing undependable respondents, 90 samples were used for final analysis. Among the 90 participants, 24 were male, and 66 were female. The respondents belong to group 1 were 32, and group 2 were 58. Particularly, the group 2 sample was collected homogenously from one MCN called Korea New Media Group. Freelance announcers whose images and videos are continuously being photographed or broadcasted through diverse channels are managed here. Detailed descriptions are summarized in Table 1.

Table1: Descriptive statistics of the respondents.

Measures		Frequency	Percent
Gender	Male	24	26.7%
	Female	66	73.3%
Age	20s	53	58.9%
	30s	37	41.1%
Highest level of	High school	-	-
education	College	5	5.6%
	University	79	87.8%
	Graduate school	6	6.7%
	Other	-	-
Group	(1) SMIs	32	35.6%
	(2) Jobs	58	64.4%

3.2. Measures

The survey was composed of 3 parts: the first part investigated individuals' social media use/activity and deepfake knowledge; the second part was about participants' evaluation of the noise-added images comparing to the original image, and the third part measured individuals' intention to use the noise. Short questions asking demographic information were included in the last part of the survey.

Part 1: Prior to the main study, this study generated questions for social media usage behavior and deepfake knowledge. The motivations for posting images and videos on social media were asked to the participants with the items borrowed from Sung et al. (2016). To identify the samples' familiarity or knowledge of the deepfake, this study asked whether they have heard about, know of, or seen any deepfake content.

Part 2: To assess the optimal noise level from the user perspective, this study adopted image

quality measurement from Noh & Har (2010) and modified it to fit the research context. A summarized explanation of the mechanism of deepfake prevention through adversarial noise was provided at the beginning of the evaluation. This study has constructed evaluation items with adjectives that can be easily understood and evaluated by anyone. Measurement items in the scale were derived from Radun et al. (2008), who developed subjective image-quality evaluations with an interpretation-based estimation of image quality. The adjectives were revised concerning Engeldrum's (2001) psychometric scaling of putting '-ness' on image quality measures. Next, in order to minimize external factors such as types of content or survey environment, this study presented three images simultaneously to minimize the impact of the content. Lastly, to report the quality of noise-added images compared to the original image, the original image was provided before every noise-added image appears. In sum, the survey of part 2 was processed as following: presenting three original images → presenting the noise-added image in level 0.009, respectively → participants measuring the quality of a noise-coated image at 0.009. The exact process was iterated with 0.03 and 0.5 noise levels. For the data analysis, this study used SPSS 25.0 for data reduction, coding, and calculating the means as the overall quality of the image. One-way analyses of variance (ANOVAs), which test the difference between two or more groups, were used as the analytic tool (Fraiman & Fraiman, 2018). To be more specific, the researcher divided the groups according to the images (e.g., group1 – image1) and the mean values for each noise level within the group (e.g., mean values of image 1's 0.009 – 0.03 – 0.5 noise level). The total mean values were calculated by averaging the values of seven measurement items. Then, this study looked at whether there are statistically significant differences in people's overall image quality judgments to 0.009 – 0.03 – 0.5 noise-added images with ANOVA.

Part 3: In order to identify individuals' intention to use the noise, this study constructed items for each hypothesis from previous literature with some minor modifications to fit the context of the study. The measures for intention, attitude, the benefit of using the noise, cost of using the noise, and cost of not using the noise, intrinsic benefit, resource safety, resource vulnerability, and work impediment were derived from Bulgurcu et al. (2010). The measures for social norm, perceived behavioral control (PBC), privacy risk, and threat severity was adopted from Taneja et al. (2014). All items were measured with a seven-point Likert scale ranging from "1=strongly disagree" to "7=strongly agree". This study employed partial least squares structural equation modeling (PLS-SEM). Using SmartPLS 3.0, it examined the validity and reliability of the measurement items and the proposed hypotheses. Wong (2013) suggested that PLS-SEM is particularly suitable when the model is in its early stage. As this study adopts the ISP compliance model by Bulgurcu et al. (2010), which is quite a lately model that needs further hypothetical testing, PLS-SEM appears to be the just method to analyze the research model.

4. Results

4.1. Results of image quality assessment

ANOVA tests for the difference between means of the individuals' evaluation on the three different noise levels are proved to be statistically significant. As shown in Table 2, the ANOVA results for image 1 show that the noise-added images are significantly different from each other according to the noise level ($F=296.165$, $p<.001$). Subsequently, a post hoc test in ANOVA by each of the noise level groups was conducted. The results revealed significant differences in image quality scores among 0.009 and 0.5 levels and 0.03 and 0.5 levels. In addition, the mean values of 0.009 and 0.03 clusters are close to four, which means that participants noticed no difference from the original. Table 3 with the results for image 2 ($F=812.670$, $p<0.001$) and Table 4 with the results for image 3 ($F=698.311$, $p<0.001$) all revealed to be statistically significant. Similar tendencies were found in post hoc tests also for images 2 and 3. While 0.009 and 0.03 showed almost no differences in quality scores, the evaluation for 0.5 was mainly different from the other two levels, in a negative direction. For instance, for image 1, while the mean value of 0.009 is 4.4557, which is very close to the 4 (absolutely original), the mean of 0.5 is 1.7443, which is profoundly negative and unlike the others. These results indicate that ordinary people do not perceive the differences between 0.009 and 0.03 noise levels, and they are both considered similar to the original image. However, people's response is different at the noise level of 0.5, which turns out to be negative.

Table 2: ANOVA table for image 1.

Dependent variable	Classification = Noise level	Mean	SD	F ratio / significance	Post-hoc test (amount of difference)
Evaluations on the noise-added image (1)	(1) 0.009	4.4557	.28213	296.165 /.000***	(1) - (3) 2.711*
	(2) 0.03	4.3871	.13720		(2) - (3) 2.643*
	(3) 0.5	1.7443	.26657		(Scheffe)

Table 3: ANOVA table for image 2.

Dependent variable	Classification = Noise level	Mean	SD	F ratio / significance	Post-hoc test (amount of difference)
Evaluations on the noise-added image (2)	(1) 0.009	4.6871	.16490	812.670 /.000***	(1) - (3) 2.97429*
	(2) 0.03	4.5600	.08165		(2) - (3) 2.84714*
	(3) 0.5	1.7129	.19805		(Scheffe)

Table 4: ANOVA table for image 3.

Dependent variable	Classification = Noise level	Mean	SD	F ratio / significance	Post-hoc test (amount of difference)
Evaluations on the noise-added image (3)	(1) 0.009	4.4671	.19024	698.311/.000***	(1) - (3) 2.95143*
	(2) 0.03	4.3071	.16909		(2) - (3) 2.79143*
	(3) 0.5	1.5157	.13439		(Scheffe)

4.2. Results of partial least squares structural equation modeling

4.2.1. Test of measurement model

The measurement model for PLS-SEM is assessed by reliability, convergent validity, and discriminant validity. As shown in Table 5, the reliability of the items was demonstrated by examining the composite reliabilities (CR), Cronbach's alpha, and the average variance extracted (AVE). All of the constructs were above 0.70 for both CR (Gefen, Straub, & Boudreau, 2000) and Cronbach's alpha (Hair et al., 2006), which is the threshold that ascertains the internal consistency. The AVE for all construct measurements was above 0.50 (Gefen et al., 2000), thereby ascertaining the convergent validity.

Following Gefen and Ridings (2003), discriminant validity was confirmed by evaluating the square root of the construct's AVE that is greater than the correlations between other constructs. Moreover, each item was highly related to its own construct, proving that outer loadings were higher than other cross-loading variables in the same row and column.

Table 5: Reliability and validity results.

Construct		AVE (threshold 0.5)	CR (threshold 0.7)	Cronbach Alpha (threshold 0.7)
Attitude	A	0.721	0.885	0.802
Benefit using the noise	BU	0.932	0.976	0.963
Cost not using the noise	CNU	0.903	0.965	0.948
Cost using the noise	CU	0.824	0.933	0.893
Intention to use	IU	0.784	0.947	0.930
Intrinsic benefit	IB	0.889	0.960	0.938
Perceived behavioral control	PBC	0.651	0.840	0.763
Privacy risk	PR	0.843	0.942	0.907
Resource distortion	RD	0.864	0.962	0.948
Resource safety	RS	0.859	0.960	0.945
Resource vulnerability	RV	0.897	0.972	0.962
Social norm	SN	0.785	0.916	0.863

Threat severity	TS	0.971	0.990	0.985
Work impediment	WI	0.738	0.918	0.880

Note: AVE = average variance extracted; CR = composite reliability

Table 6: Loadings and cross loadings.

	A	BU	CNU	CU	IB	IU	PBC	PR	RD	RS	RV	SN	TS	WI
A1	**0.735**	0.467	0.333	-0.252	0.403	0.443	0.017	0.200	-0.091	0.425	0.372	0.453	0.338	-0.303
A2	**0.908**	0.609	0.349	-0.143	0.639	0.575	0.287	0.260	-0.030	0.563	0.387	0.508	0.346	-0.177
A3	**0.893**	0.570	0.339	-0.020	0.645	0.579	0.140	0.315	-0.056	0.372	0.413	0.595	0.322	-0.153
BU1	0.622	**0.964**	0.367	0.024	0.701	0.646	0.248	0.351	0.074	0.721	0.531	0.479	0.498	-0.176
BU2	0.616	**0.977**	0.269	0.023	0.724	0.623	0.250	0.316	0.058	0.715	0.535	0.497	0.461	-0.159
BU3	0.643	**0.954**	0.360	0.063	0.723	0.584	0.158	0.297	0.028	0.646	0.511	0.514	0.443	-0.110
CNU1	0.453	0.388	**0.956**	-0.175	0.268	0.602	0.169	0.402	-0.034	0.297	0.438	0.437	0.437	-0.200
CNU2	0.382	0.313	**0.959**	-0.113	0.200	0.550	0.187	0.277	-0.045	0.212	0.324	0.327	0.343	-0.100
CNU3	0.249	0.239	**0.936**	-0.093	0.079	0.428	0.139	0.200	-0.115	0.139	0.277	0.245	0.261	-0.051
CU1	-0.182	-0.011	-0.160	**0.955**	-0.043	-0.006	-0.111	-0.161	0.414	-0.252	-0.173	-0.334	-0.084	0.714
CU2	-0.160	0.066	-0.160	**0.922**	-0.076	0.091	-0.095	-0.174	0.513	-0.208	-0.002	-0.241	0.052	0.727
CU3	-0.062	0.052	-0.045	**0.842**	0.078	0.112	-0.113	-0.065	0.336	-0.225	-0.018	-0.129	0.025	0.539
IB1	0.614	0.664	0.175	-0.069	**0.940**	0.515	0.206	0.396	0.044	0.513	0.363	0.539	0.354	-0.344
IB2	0.624	0.739	0.180	-0.015	**0.943**	0.475	0.187	0.364	-0.091	0.532	0.393	0.505	0.388	-0.304
IB3	0.660	0.692	0.240	0.010	**0.946**	0.558	0.237	0.429	-0.005	0.486	0.443	0.572	0.405	-0.299
IU1	0.532	0.475	0.537	0.001	0.429	**0.821**	0.144	0.253	0.243	0.433	0.311	0.392	0.287	-0.055
IU2	0.613	0.635	0.544	0.074	0.518	**0.933**	0.266	0.368	0.178	0.491	0.455	0.492	0.417	-0.005
IU3	0.598	0.613	0.574	0.060	0.523	**0.942**	0.328	0.341	0.200	0.513	0.407	0.437	0.420	0.007
IU4	0.587	0.591	0.495	0.105	0.481	**0.940**	0.220	0.305	0.230	0.469	0.406	0.400	0.421	0.054
IU5	0.436	0.502	0.343	0.052	0.470	**0.778**	0.105	0.284	0.366	0.428	0.294	0.319	0.238	-0.035
PBC1	-0.028	-0.006	0.254	-0.290	-0.056	0.035	**0.511**	0.086	-0.198	0.141	0.136	0.131	0.209	-0.101
PBC2	0.220	0.254	0.187	-0.090	0.228	0.289	**0.965**	0.287	-0.028	0.416	0.104	0.088	0.191	-0.098
PBC3	0.108	0.158	0.080	-0.088	0.208	0.146	**0.872**	0.317	0.008	0.261	0.165	0.076	0.159	-0.105
PR1	0.228	0.253	0.338	-0.153	0.348	0.303	0.291	**0.907**	-0.077	0.263	0.361	0.411	0.265	-0.295
PR2	0.291	0.308	0.276	-0.102	0.378	0.274	0.273	**0.929**	0.044	0.324	0.337	0.418	0.225	-0.198
PR3	0.333	0.363	0.282	-0.164	0.433	0.394	0.293	**0.919**	0.017	0.401	0.401	0.414	0.289	-0.260
RD1	0.009	0.163	-0.030	0.492	0.057	0.296	-0.029	0.030	**0.922**	-0.026	-0.036	-0.123	0.018	0.425
RD2	-0.091	-0.007	-0.068	0.369	-0.075	0.219	-0.096	-0.021	**0.944**	-0.135	-0.071	-0.127	-0.043	0.410
RD3	-0.060	0.048	-0.039	0.444	0.036	0.251	0.025	0.024	**0.929**	-0.094	-0.058	-0.080	-0.011	0.383
RD4	-0.121	-0.025	-0.091	0.429	-0.118	0.198	-0.027	-0.079	**0.923**	-0.136	-0.057	-0.126	-0.036	0.491
RS1	0.586	0.691	0.179	-0.251	0.613	0.458	0.338	0.440	-0.120	**0.910**	0.408	0.385	0.299	-0.351
RS2	0.477	0.729	0.203	-0.217	0.515	0.483	0.410	0.356	-0.061	**0.948**	0.347	0.345	0.278	-0.267
RS3	0.467	0.652	0.272	-0.212	0.426	0.515	0.375	0.276	-0.127	**0.945**	0.268	0.297	0.286	-0.231
RS4	0.451	0.578	0.254	-0.252	0.439	0.508	0.293	0.229	-0.065	**0.903**	0.233	0.265	0.296	-0.301
RV1	0.407	0.470	0.332	-0.146	0.369	0.353	0.194	0.407	-0.149	0.340	**0.949**	0.421	0.773	-0.253

RV2	0.456	0.550	0.341	-0.066	0.422	0.413	0.131	0.372	-0.074	0.361	**0.959**	0.468	0.817	-0.177
RV3	0.487	0.551	0.353	-0.053	0.461	0.448	0.142	0.387	-0.036	0.351	**0.964**	0.434	0.826	-0.195
RV4	0.393	0.492	0.402	-0.028	0.357	0.404	0.072	0.351	0.019	0.259	**0.916**	0.345	0.678	-0.122
SN1	0.593	0.469	0.357	-0.280	0.536	0.467	0.060	0.435	-0.109	0.335	0.442	**0.965**	0.338	-0.357
SN2	0.558	0.548	0.375	-0.168	0.466	0.447	0.177	0.440	-0.046	0.410	0.449	**0.894**	0.319	-0.259
SN3	0.463	0.316	0.233	-0.287	0.541	0.294	-0.007	0.302	-0.209	0.139	0.226	**0.790**	0.149	-0.354
TS1	0.373	0.431	0.378	-0.015	0.374	0.374	0.208	0.265	-0.014	0.283	0.788	0.277	**0.985**	-0.148
TS2	0.418	0.512	0.373	-0.006	0.423	0.430	0.179	0.312	-0.017	0.328	0.828	0.355	**0.983**	-0.102
TS3	0.373	0.489	0.378	0.007	0.404	0.413	0.210	0.262	-0.019	0.311	0.790	0.309	**0.989**	-0.115
WI1	-0.254	-0.168	-0.104	0.596	-0.331	0.015	-0.176	-0.237	0.424	-0.311	-0.148	-0.310	-0.068	**0.910**
WI2	-0.333	-0.207	-0.094	0.573	-0.390	-0.030	-0.122	-0.249	0.397	-0.366	-0.205	-0.333	-0.171	**0.891**
WI3	-0.351	-0.233	-0.136	0.549	-0.464	-0.104	-0.107	-0.274	0.349	-0.324	-0.231	-0.347	-0.185	**0.893**
WI4	0.041	0.032	-0.129	0.741	-0.029	0.076	0.001	-0.191	0.386	-0.100	-0.095	-0.241	-0.024	**0.729**

Table 7: Correlation matrix.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	11	12	13	14
A	**0.849**													
BU	0.650	**0.965**												
CNU	0.400	0.344	**0.950**											
CU	-0.066	0.038	-0.141	**0.908**										
IU	0.630	0.640	0.570	0.068	**0.885**									
IB	0.671	0.742	0.210	-0.250	0.546	**0.943**								
PBC	0.185	0.227	0.176	-0.116	0.250	0.222	**0.807**							
PR	0.307	0.333	0.328	-0.154	0.353	0.419	0.312	**0.918**						
RD	-0.066	0.055	-0.060	0.471	0.263	-0.021	-0.033	-0.010	**0.930**					
RS	0.536	0.719	0.242	-0.250	0.528	0.542	0.385	0.356	-0.101	**0.927**				
RV	0.459	0.545	0.380	-0.074	0.429	0.424	0.139	0.400	-0.058	0.343	**0.947**			
SN	0.612	0.515	0.372	-0.268	0.466	0.570	0.097	0.451	-0.123	0.351	0.437	**0.886**		
TS	0.393	0.484	0.382	-0.005	0.412	0.406	0.202	0.284	-0.017	0.312	0.814	0.318	**0.986**	
WI	-0.241	-0.154	-0.137	0.737	-0.004	-0.334	-0.111	-0.276	0.460	-0.310	-0.193	-0.357	-0.123	**0.859**

4.2.2. Results of hypothesis testing

A bootstrapping method with 5000 bootstrap subsamples was conducted to estimate the significance of the structural model (Wong, 2013). The research model of this study explained 42.7% of the variance for intention, 48.0% of the variance for attitude, 69.3% for the benefit of using privacy controls, 20.1% for the cost of not using privacy controls, and 56.5% for the cost of using privacy controls. Figure 2 summarizes the results of the structural model analysis with r-square, path coefficients, and p-values.

The relationship between attitude and intention was found to be significant ($\beta = 0.516$, $t = 4.561$, $p < .001$), ensuring H1. However, there was no significant effect of social norm and perceived behavioral control on intention, implying that H2 and H3 were not supported. A statistically significant positive effect was found between benefit of using the noise and attitude ($\beta = 0.597$, $t = 8.120$, $p < .001$) and cost of not using the noise and attitude ($\beta = 0.171$, $t = 2.274$, $p < .05$), verifying H4 and H5. On the other hand, while the cost of using the noise showed a negative effect on attitude, this effect was a negligible one, not supporting H6. Intrinsic benefit ($\beta = 0.502$, $t = 5.108$, $p < .001$) and resource safety ($\beta = 0.445$, $t = 4.391$, $p < .001$) were found to be significant in increasing the benefit of using the noise, which were consistent with H7 and H8. In contrast, threat severity and resource vulnerability did not significantly affect the cost of not using the noise, showing inconsistent results with H9 and H10. Privacy risk was only a significant predictor for the cost of not using the noise ($\beta = 0.232$, $t = 2.704$, $p < .05$), which verified H11. The relationship between work impediment and cost of using the noise was found to be significant ($\beta = 0.655$, $t = 9.190$, $p < .001$), as well as the relationship between resource distortion and cost of using the noise ($\beta = 0.176$, $t = 2.130$, $p < .05$). Thus, the results supported H12 and H13. Table 8 summarizes the results of path coefficient analysis.

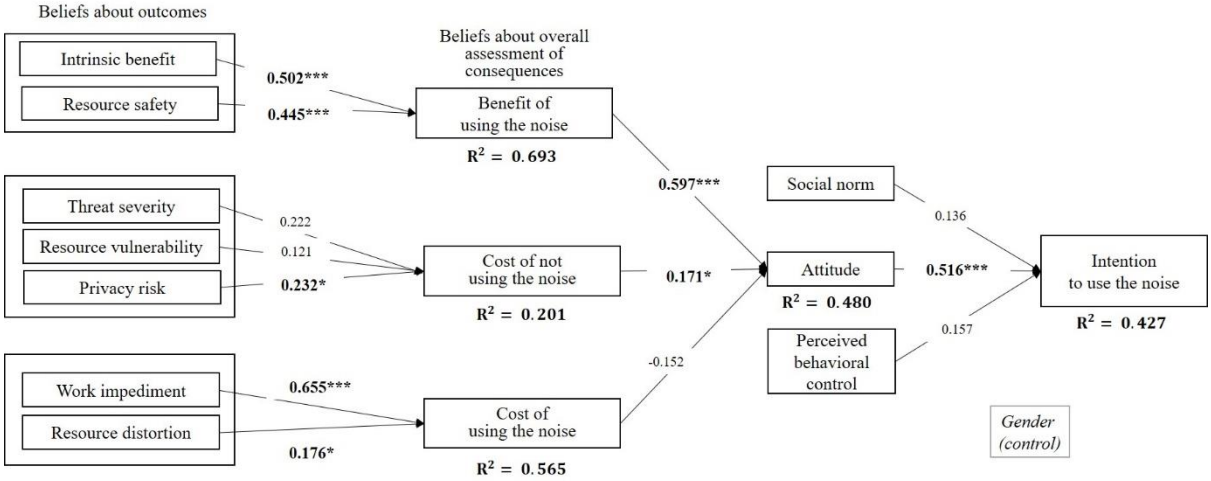

Figure 3: Results of the research model.

Table 8: Path coefficient analysis.

	Hypothesis	β	t	Result
H1	Attitude → Intention	0.516	4.561***	Accepted
H2	Social norm → intention	0.136	1.167	Rejected
H3	Perceived control → intention	0.157	1.664	Rejected
H4	Benefit of using Noise → attitude	0.597	8.120***	Accepted
H5	Cost of not using Noise → attitude	0.171	2.274*	Accepted

H6	Cost of using Noise → attitude	-0.152	1.897	Rejected
H7	Intrinsic benefit → Benefit of using Noise	0.502	5.108***	Accepted
H8	Resource safety → Benefit of using Noise	0.445	4.391***	Accepted
H9	Threat severity → Cost of not using Noise	0.222	1.239	Rejected
H10	Resource vulnerability → Cost of not using Noise	0.121	0.449	Rejected
H11	Privacy risk → Cost of not using Noise	0.232	2.074*	Accepted
H12	Work impediment → Cost of using noise	0.655	9.190***	Accepted
H13	Resource distortion → Cost of using Noise	0.176	2.130*	Accepted

5. Discussions and Conclusions

5.1. Discussions

This paper examined individuals' responses to the noise added to the original image, analyzing the proper algorithm and user-perceived noise level and the intention to use the noise technology for personal image protection.

In study 1, this study conducted experiments with StarGAN, a type of generative model, and PGD, an adversarial attack algorithm, to identify an appropriate noise level to reduce the performance of deepfake generation algorithms. The results show that deepfake starts malfunctioning from 0.009 noise level. Higher noise levels increase the degree of protection by keeping deepfake away from transforming the images. In other words, in terms of technical defense, it is recommended to raise the noise level higher to prevent deepfakes flawlessly. However, the findings suggest that a higher noise level brings about more significant damage to the original image. If the noise increases over a critical point, people might express uncomfortable feelings to the image even with the naked eye. In sum, study 1 not only discovered the starting point of deepfake malfunctioning but also enabled further research on a user-based investigation to find acceptable noise levels.

The second study analyzed the individuals' response to the noise-added images compared to the original one. In order to develop comprehensive image quality evaluation criteria, this study adopted a measurement scale from Noh & Har (2010) and Radun et al. (2008). As a result of the evaluation analysis, the study demonstrated that people found little difference from the original when the noise was at 0.009 and 0.03. Almost no discrepancy was found between 0.009 and 0.03. However, in the case of 0.5, the quality of the image was turned out to be significantly different from the other two noise levels in a negative fashion. In addition, the degree to which people perceive differences from the original image was high. These findings demonstrate that the damage to the original image with about 0.5 noise is not acceptable to the user's eye. Instead, because the participants have hardly detected any

difference with the naked eye at the 0.009 or 0.03 noise level, they seem just adequate to maintain image quality while effectively defending deepfake attacks.

After discovering the right level of noise from the user perspective, this study explored individuals' intention to use the noise as a privacy-keeping device. Even the study built a model that could preemptively avert deepfake attacks with acceptable noise levels; it will become of no use if the user finds no intention to use the noise. As such, this part of the study examined what kind of determinants can predict people's adoption of noise. Accordingly, this study adopted Bulgurcu's ISP compliance model framework in the context of deepfake prevention to measure and test the relationship between individuals' beliefs about outcomes and assessment of consequences of using or not using the noise. Although Bulgurcu's study focused on individuals' either mandatory or voluntary behavior about already existing policy under organization, this study investigated only the voluntary intent of individuals.

The results indicate that intention to use the noise is highly influenced by individuals' attitudes on the noise, as scholars predicted. Individuals' beliefs that influence attitudes were the benefits of using and the cost of not using the noise. On the contrary to the study's hypothesis based on TPB, social norm and perceived behavior did not have a significant effect on the intention to use the noise. Social norm was yet to be a controversial factor with mixed results proposed in the previous literature; for example, Venkatesh et al. (2003) discovered that social norms are only influential in a mandatory setting and more salient to older workers. This study assumed a situation of entirely voluntary choice to use new technologies on young adults aged between the 20s and 30s. Thus, this outcome was not unconvincing. Further, PBC was not related to the intention to use the noise. In the deepfake context, it is physically impossible for individuals to control any privacy control mechanism because there has never been any technology before. Thus, the survey participants had to predict their expected ability to use the noise on the social media platform. In situations where and how this technology may be distributed or constructed are still in question, measuring the PBC item may have been difficult for the participants.

The results also show that the benefit of using the noise is most significantly related to attitude on the noise followed by the cost of not using it. However, no relationship could be found between the cost of using the noise and attitude, which is consistent with previous literature (Taneja et al., 2014). This suggests that it is more important to emphasize the advantages of using the noise and the disadvantages of not using the noise, explaining how the benefit of the technology can outweigh any possible cost. Interestingly, although statistically insignificant, a negative relationship was captured between the cost of using the noise and attitude, as we expected. In line with the extant studies, users may feel annoyance or confusion in using the new technology and even feel the loss of enjoyment on social media usage (Brandtzæg et al., 2010; Thambusamy et al., 2010). Thus, the findings reveal that it is vital to reduce the burden of using new technologies. If a noise-adding solution were to be

commercialized in the future, it would be better to make a straightforward, not complexed setting.

Furthermore, the study's findings on each belief about outcomes on individuals' cost and benefit assessment are also worth noting. The results affirm that intrinsic benefit and resource safety both exert a powerful impact on the benefit of using the noise. Individuals will acknowledge more advantages when they feel comfortable and satisfied about protecting their resources from misuse. In other words, the results tell that the usability of the noise depends upon the belief that their images and videos are safe and protective from any deepfake abusing threat.

While threat severity and resource vulnerability did not significantly influence the cost of not using the noise, privacy risk had a significant positive effect on not using the noise. This suggests that individuals worry about possible privacy-related risks like deepfake sex crime or ruined reputation that may result from not using the deepfake prevention technology. Nevertheless, in this study's context, the fear of privacy invasion and the perception about possible ill-usage of personal resources may have no relationship with the beliefs on the unfavorable consequence of not using the noise. First, the main participants of the study were SMIs, the ones who engage in various activities on social media although knowing these problems. Second, in a similar vein to the early discussion, since people have never seen or heard about this technology, they may not be able to link the threats or vulnerability with the cost of not using it.

The findings demonstrate that work impediment and resource distortion are essential determinants for the cost of using the noise. Individuals are found to post their selfies on SNSs for many reasons: attention-seeking, communication, archiving, or entertainment (Sung et al., 2015). If individuals have to use the noise that is likely to hinder these drivers from posting images on SNSs, they may perceive the noise as a cost. This may be why the higher the work impediment predicted, the higher the cost of using the noise. In terms of resource distortion, if the noise distorts the color, sharpness, or clarity of the original image and its overall preference, individuals who value greatly uploading images and videos on social media platforms may perceive the noise as the noise potential cost.

5.2. Implications

This study also provides distinct insights for practitioners, especially potential customers of this technology, the platform providers. The results of studies 1 and 2 suggest that although maximally increasing the noise would be better for a perfect defense to the deepfake, minimally 0.009 up to 0.03 is found to be the appropriate level for general people to use the technology without any inconvenience. Thus, this study suggests that platform providers willing to embrace the noise should consider providing privacy control within these levels. In such a way, successful utilization of the noise as preemptive

deepfake blocking techniques is expected to solve the potential socio-political risks of deepfakes.

The results for study 3 indicate that the intention to use noise derives from a positive attitude toward the noise. That attitude is formed from the evaluations of various advantages that individuals can get from using the noise and disadvantages that individuals can lose without using the noise. The study highlights the necessity to emphasize the benefits of using the noise, such as individuals' inner satisfaction and safety of resources online. Hence, it would be encouraged for the practitioners to emphasize users' advantages as the primary value of the prevention method. Furthermore, the results pinpoint that people's fear of not using the noise has significant relevance with threats to possible privacy risks. In this respect, convincing users that the noise is a guaranteed method to prevent the risks of AI, such as malicious image manipulation, would be the most efficient to promote the technology. In addition, distortion of images and work impediments are turned out to be the components that have a significant impact on the cost of using the noise. Accordingly, minimizing the distortion of the original image and the cost of time, money, and effort looks profoundly important in enhancing the usage intention of the noise. In sum, it is recommended for platform operators to appeal that the noise is the best way to protect private data and the simplest.

This paper also has implications in the social aspect. The survey results showed that people failed to link the severity of the potential deepfake-associated privacy risks to the cost of not using the noise. This finding likely implies that digital native generations and social influencers are expected to lack awareness of privacy-related problems on social media. This phenomenon could have occurred as posting content on personal social accounts has become the everyday life of SMIs and young people. Further, individuals may have perceived that deepfake risks as someone else's story since most of the deepfake abuse reported in the news or social media are mainly targeted to well-known politicians and celebrities. However, a series of criminal issues in South Korea and numerous sexually demeaning hashtags on Twitter with deepfake technology clearly show that the deepfake issue is no longer someone else's. In this regard, it is necessary for diverse stakeholders to explain the perceived damage or risk of deepfake accurately: at the micro-level, social media e-WOM through the relationship between followers/friends or content, at the meso-level, schools or colleges as proper educating herb, and at the macro-level, a governmental organization from local to central by setting up a watch group, can all play this role. Although this study focused on proactive deepfake prevention, assuming the circumstances in which these technologies will be used, this field of study is still rudimentary and complexed. It is even hard to tell when the time for general users would be to use this technology on media platforms. Thus, governmental level support for deepfake prevention and detection technology and legislation that effectively punish offenders and protect victims is yet to be of great importance.

5.3. Limitations and future research

This study has not without limitations. To begin with, the sample used in this study are social media influencers or freelance announcers, the results may have a limited generalizability to general population. Second, as the survey was conducted an online, and the data was collected through snowballing method, the number of samples is quite insufficient, and some demographic bias could have occurred. Third, although Bulcurgu's ISP compliance model meaningfully explains individual's intention to use, other potential factors that could explain were not taken into consideration, which should be further examined in future research. A possible research direction would be to examine other groups of people, with larger amount of data, and other possible variables that can explain intention to use the noise.

References

[1] Ajder, H., Patrini, G., Cavalli, F., & Cullen, L. (2019). The state of deepfakes: landscape, threats, and impact. Amsterdam: Deeptrace. http://regmedia.co.uk/2019/10/08/deepfake_report.pdf.

[2] Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, *50*(2), 179-211.

[3] Bates, M. E. (2018). Say what? "deepfakes" are deeply concerning. *Online Searcher, 42*(4), 64.

[4] Bulgurcu, B., Cavusoglu, H., & Benbasat, I. (2010). Information security policy compliance: an empirical study of rationality-based beliefs and information security awareness. *MIS quarterly*, 523-548.

[5] Chesney, B., & Citron, D. (2019). Deep fakes: a looming challenge for privacy, democracy, and national security. *California Law Review, 107*, 1753.

[6] Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., & Bharath, A. A. (2018). Generative adversarial networks: An overview. *IEEE Signal Processing Magazine*, *35*(1), 53-65.

[7] Debatin, B., Lovejoy, J. P., Horn, A. K., & Hughes, B. N. (2009). Facebook and online privacy: Attitudes, behaviors, and unintended consequences. *Journal of Computer-Mediated Communication*, *15*(1), 83-108.

[8] Engeldrum, P. G. (2001). Psychometric scaling: avoiding the pitfalls and hazards. *PICS,* 101-107.

[9] Fishbein, M., & Ajzen, I. (1977). *Belief, attitude, intention, and behavior: An introduction to theory and research*. ding, MA: Addison-Wesley

[10] Fraiman, D., & Fraiman, R. (2018). An ANOVA approach for statistical comparisons of brain networks. *Scientific Reports*, *8*(1), 1-14.

[11] Gefen, D., & Ridings, C. M. (2003). IT acceptance: managing user—IT group boundaries. *ACM SIGMIS Database: the DATABASE for Advances in Information Systems*, *34*(3), 25-40.

[12] ISO 20462 (2004). Photography - Psychophysical experimental methods to estimate image quality, *International Organization for Standardization*.

[13] Johnston, A. C., & Warkentin, M. (2010). Fear appeals and information security behaviors: An empirical study. *MIS Quarterly, 34*(3), 549–566.

[14] Keelan, B. (2002). *Handbook of image quality: characterization and prediction*. CRC Press.

[15] Li, H., Sarathy, R., & Xu, H. (2010). Understanding situational online information disclosure as a privacy calculus. *Journal of Computer Information Systems*, *51*(1), 62-71.

[16] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2017). *Towards deep learning models resistant to adversarial attacks*. https://arXiv:1706.06083.

[17] Mathieson, K. (1991). Predicting user intentions: comparing the technology acceptance model with the theory of planned behavior. *Information Systems Research*, *2*(3), 173-191.

[18] Myyry, L., Siponen, M., Pahnila, S., Vartiainen, T., & Vance, A. (2009). What levels of moral reasoning and values explain adherence to information security rules? An empirical study. *European Journal of Information Systems*, *18*(2), 126-139.

[19] Noh, Y. S., & Har, D. H. (2010). Methods of subjective image quality evaluation in pictorial images. *Journal of the Korea Contents Association*, 10(8), 186-197

[20] Pahnila, S., Siponen, M., & Mahmood, A. (2007). Employees' behavior towards IS security policy compliance, in *Proceedings of the 40fh Hawaii International Conference on System Sciences,* Los Alamitos, CA: IEEE Computer Society Press, 156-166.

[21] Park, H. J., & Har, D. H. (2016). A study on automatic measurement program for emotional preference of portraits. *The Korean Society of Science & Art*, 26, 165-177.

[22] Paternoster, R., & Pogarsky, G. (2009). Rational choice, agency and thoughtfully reflective decision making: The short and long-term consequences of making good choices. *Journal of Quantitative Criminology*, *25*(2), 103-127.

[23] Pedersen, M., Bonnier, N., Hardeberg, J. Y., & Albregtsen, F. (2009). Attributes of a new image quality model for color prints. In *Color and Imaging Conference* (1), 204-209.

[24] Rogers, R. W. (1975). A protection motivation theory of fear appeals and attitude change1. *The Journal of Psychology*, *91*(1), 93-114.

[25] Szegedy, C., Toshev, A., & Erhan, D. (2013). *Deep neural networks for object detection*.

[26] Taneja, A., Vitrano, J., & Gengo, N. J. (2014). Rationality-based beliefs affecting individual's attitude and intention to use privacy controls on Facebook: An empirical investigation. *Computers in Human Behavior*, *38*, 159-173.

[27] Taylor, S., & Todd, P. (1995). Decomposition and crossover effects in the theory of planned behavior: A study of consumer adoption intentions. *International Journal of Research in Marketing*, *12*(2), 137-155.

[28] Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS quarterly*, 425-478.

[29] Viane, I., Crombez, G., Eccleston, C., Devulder, J., & De Corte, W. (2004). Acceptance of the unpleasant reality of chronic pain: effects upon attention to pain and engagement with daily activities. *Pain*, *112*(3), 282-288.

[30] Wigfield, A. (1994). Expectancy-value theory of achievement motivation: A developmental perspective. *Educational Psychology Review*, *6*(1), 49-78.

[31] Willison, R. (2006). Understanding the perpetration of employee computer crime in the organisational context. *Information and Organization*, *16*(4), 304-324.

[32] Westerlund, M. (2019). The emergence of deepfake technology: a review. *Technology Innovation Management Review*, *9*(11), 39–52.

[33] Wong, K. K. K. (2013). Partial least squares structural equation modeling (PLS-SEM) techniques using SmartPLS. *Marketing Bulletin*, *24*(1), 1-32.

[34] Xu, H., Dinev, T., Smith, H. J., & Hart, P. (2008). Examining the formation of individual's privacy concerns: Toward an integrative view. *In Proceedings of the International Conference on Information Systems*. Paris, France.

[35] Yendrikhovskij, S. N., Blommaert, F. J. J., & De Ridder, H. (1999). Color reproduction and the naturalness constraint. *Color Research & Application: Endorsed by Inter-Society Color Council, The Colour Group (Great Britain), Canadian Society for Color, Color Science Association of Japan, Dutch Society for the Study of Color, The Swedish Colour Centre Foundation, Colour Society of Australia, Centre Français de la Couleur*, *24*(1), 52-67.