인지과학, 제30권 제3호
Korean Journal of Cognitive Science
2019, Vol. 30, No. 3, 157～174
http://dx.doi.org/10.19066/cogsci.2019.30.3.3

Combating Identity Threat of Machine: The effect of group-affirmation on humans' intellectual performance loss*

Young-Jae Cha[1] Sojung Baek[1] Hyung-Suk Lee[2] Jonghoon Bae[3]
Jongho Lee[4] Sang-Hun Lee[5] Gunhee Kim[6] Dayk Jang[7]†

[1]Interdisciplinary Program in Cognitive Science [2]Program in History and Philosophy of Science
[3]Graduate School of Business [4]Department of Electrical and Computer Engineering
[5]Department of Brain and Cognitive Science [6]Department of Computer Science and Engineering
[7]College of Liberal Studies
All the Authors are affiliated in Seoul National University, Republic of Korea

Motivation of human individuals to perform on intellectual tasks can be hampered by identity threat from intellectual machines. A laboratory experiment examined whether individuals' performance loss on intellectual tasks appears under human identity threat. Additionally, by affirming alternative attributes of human identity, researchers checked whether group-affirmation alleviate the performance loss on intellectual tasks. This research predicted that under high social identity threat, individuals' performance loss on the intellectual tasks would be moderated by valuing alternative attributes of human identity. Experiment shows that when social identity threat is increased, human individuals affirmed alternative human attributes show higher performance on intellectual tasks than individuals non-affirmed. This effect of human-group level affirmation on performance loss did not appear in the condition of low social identity threat. Theoretical and practical implications were discussed.

Key words : Social Identity, Identity Threat, Performance Loss, Human-Machine Competition, Group Affirmation, Intergroup Relations

---

* This work was supported by the Seoul National University Research Grant in 2016.

† Corresponding author: Dayk Jang, College of Liberal Studies; Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul Republic of Korea 08826
Research areas: Evolutionary psychology, Philosophy of science
E-mail: djang@snu.ac.kr

Threats to human identity lie at the heart of the imitation game, a comparative test for machine intelligence (Turing, 1950). Machines have been developed to imitate and prove their human-level intellectual skills on games such as Chess or Go, which has been used as comparative contexts to measure the human intellectual capability against machine (for a historical review, see Table 1). Such constant challenges to the unique status of human intelligence have blurred the human-machine boundaries. Studies show that intellectual ability of machines superior to that of humans jeopardizes not only occupational security (e.g. Pew Research Center, 2014; British Science Association, 2016; Eurobarometer, 2012, Yeo, 2017), but also the certainty of human identity (e.g., Eyssel & Kuchenbrandt, 2012; Ferrari, Paladino & Jetten, 2016; Yogeeswaran et al., 2016). Despite increasing concerns for human underperformance, there remains a dearth of research on how people manage the threat from the superiority of machine intelligence in certain intellectual dimensions.

〈Table 1〉 A Brief History of Imitation Games between Humans and Machines

Years	Games	Results
1997	Chess	Deep Blue defeated world champion
1997	Othello	Logistello defeated world champion
2007	Scrabble	Quackle defeated the former world champion
2010	Shogi	Akara defeated shogi champion
2011	Quiz	Watson won Jeopardy! defeating former winners
2016	Go	AlphaGo defeated Korean champion
2016	Go	AlphaGo Master won 60 online games against professional Go players
2017	Go	AlphaGo Master defeated world champion
2017	Poker	Libratus defeated top-class poker players
2018	Reading Comprehension	SLQA+ and R-NET+ outscored humans in the Stanford Question Answering Dataset (SQuAD)

This research aims to reveal how people buffer the identity threat of machine intelligence. Based on the social identity perspective (SIP; Tajfel & Turner, 1979; Tajfel & Turner, 1986; Terry & Hogg, 1996; Turner & Reynolds, 2011; Hogg, Abrams, Otten, and Hinkle, 2004), we predict that people will show performance loss when they face superior abilities of machines. Specifically, people exposed to machines' superiority in intellectual dimensions will be less motivated to engage in a task within

the intellectual dimensions. To prevent the reduction of individual performance, group affirmation manipulation that puts more importance on alternative dimensions of human identity will be used.

Lines of research have shown that people lose their motivation when they engage in tasks related to the threatened dimensions of human identity. As people's self-esteem is associated with the superiority of ingroup dimensions they identify themselves with (Tajfel & Wilkes, 1963; Branscombe, Ellemers & Doosje, 1999), threats to the superiority of such dimensions may lead them to cognitively devalue and selectively disidentify themselves from the status-relevant outgroup dimensions (Crocker, Major, & Steele, 1998; Major, Spencer, Schmader, Wolfe, & Crocker, 1998). That is, people lose their motivation and disengage from the tasks that threaten their intellectual superiority over machines (Becker, 2012; Crocker et al., 1998; Derks, van Laar, & Ellemers, 2007; Osborne, 1997; Schmader, Major, Eccleston, & McCoy, 2001). Consequently, this harms an individual's performance on intellectual tasks (Baumeister & Jones, 1978; Spencer, Steele, & Quinn 1999). Therefore, it can be predicted that an individuals' perception of machine's superiority in rational intelligence will degrade his or her performance on the rationality-related tasks.

However, affirming group identity may reduce performance-loss (Derks et al., 2007). Given machine intelligence may threaten individual's human-level identity, restoring the integrity of their social identities should be effective (Sherman, Kinias, Major, Kim, & Prenovost, 2007). Unlike a self-affirmation restores one's self-integrity, a group affirmation restores an important group value (Steele, 1988). According to the research on group-affirmation literature (Derks, Scheepers, Van Laar, & Ellemers, 2010; Guun & Wilson, 2011; Sherman & Cohen, 2006), performance loss in a threatened dimension of an individual or a group can be mitigated by focusing on alternative positive dimensions of comparison. For example, Derks et al. (2007) show that people can increase their performance and motivation in the threatened dimension by valuing an ingroup dimension. Thus, we will check the preventive role of affirmation manipulation against performance loss in rationality tasks by affirming human identity (i.e., group affirmation).

To test the two hypotheses, we manipulated identity threat and examined whether the threat decreases individual's performance on intellectual tasks (performance-loss hypothesis). Secondly, we orthogonally manipulated group affirmation and checked whether the affirmation moderated the relationship between threat and performance (affirmation-as-buffer hypothesis).

Method

Participants

210 Korean undergraduate students (93 females, Mage = 22.85, SD = 2.3) enrolled in an offline study for the exchange of 10,000 Korean Republic won. Among 210 participants, 108 participants were recruited from a long-term participant pool of the Culture-Brain Dynamics Transdisciplinary Research Center at Seoul National University. The survey was administered in full compliance with the safety guidelines, as approved by the Institutional Review Board at Seoul National University.

Procedure

One to four individuals participated per session and were directed to individual seats separated by partition walls. At the beginning of the experiment, participants were informed that they would take part in a research project observing people's perception of the intergroup competition between humans and machines. It was further informed that the experiment consists of two parts; first to report their subjective thoughts and perceptions over given information related to human-machine competition, and second to engage in a task competing with machines. In the first part, two pieces of information were given as manipulating materials: threat and affirmation manipulation.

The first piece of information was threat manipulation. Participants were randomly assigned to high vs. low threat conditions. Depending on the conditions, different news articles were given to participants to be read. Articles depicting human-machine relations shared the same title, "Humans vs. Machines: competitions on human-machine rational intelligence" but differed in their contents regarding expectations on the level of threat as described below. The second piece of information was orthogonally given to participants. Participants were randomly assigned to affirmation vs. control conditions. Participants were presented with data introducing survey results made by an expert group. Survey results were differed depending on the conditions, and participants were asked to report their own opinions regarding given the survey result.

After responding several follow-up questions, participants were introduced to participate in a task of competing with machines. The task was allegedly represented as under a developmental phase in an AI institute at their university. After the introduction given by the instructor, they performed a

logical task consists of 30 questions. Being finished with the task, they reported demographics, took suspicion check, and then were debriefed.

Materials

Threat Manipulation

Two different news articles were used to manipulate the perception of the threat. This material was made in the form of a newspaper article, which provided information to induce a highly threatening (or lowly threatening) perception on differences in the rational thinking abilities between human and machine caused by machine development.

Affirmation Manipulation

Researchers have manipulated group affirmation by providing positive feedback about their group performance (e.g., Derks et al., 2010). In this study, we provided positive feedback on a threat-unrelated performance dimension. An ostensibly reputable survey result showed experts' evaluation on human cognitive openness. Participants under the affirmation condition were given experts' positive evaluations on cognitive openness. Participants were asked to elaborate his or her thoughts on the survey results from experts. Furthermore, to boost the affirmation effect, participants were instructed to freely describe the reasons of human superiority on cognitive openness. In contrast, participants under the control condition were given experts' relatively negative evaluations on cognitive openness.

Identity Threat Measurement

To measure the perceived identity threat directly, we led participants to answer to the question, "In the material you read, did the machine seem to threaten human rationality?" (6-point scale ranging from (1) not at all to (6) extremely).

Rationality-task Performance

Participants were deceived as being randomly assigned to a rationality task. The researchers explained that the results of the participants' task performance will be compared with those of A.I, which is being developed by the Institute of Artificial Intelligence at their university. The task consisted of a total of 30 logical problems and of judging whether it was logically valid to derive B from A: A) Only residents of the city can run for the mayor. B) Any resident of the city may run for the mayor. The number of correct answers was calculated as the rational task performance score. There was no time limit for this task, and participants were informed they could quit the task whenever they want.

Results

〈Table 2〉 Means, SD, and correlations

	Mean	SD	*1*	*2*	*3*	*4*
1. Threat Manipulation	.49	.50	1.00			
2. Affirmation Manipulation	.53	.50	-.07	1.00		
3. Identity Threat	4.99	1.16	-.03	.04	1.00	
4. Performance	24.17	3.05	-.02	.17*	.01	1.00

*$p < .05$

Manipulation Check

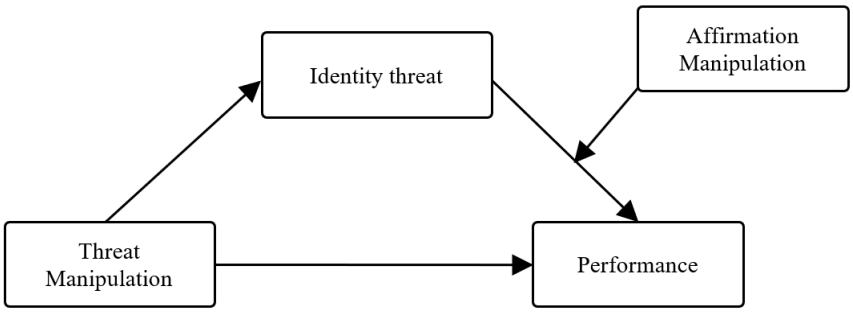
Responses to the questionnaire item asking, "It is unlikely for the advantage of artificial intelligence in rational intelligence to be easily reversed," differed between participants under the high threat condition (M = 4.67, SD = 1.11) and the low threat condition (M = 3.69, SD = 1.22) in a statistically significant manner (F(1,199) = 35.65, p < .001, and η2p = .152). The affirmation manipulation did not affect the threat perception (F(1,199) = 2.76, and p = .153). No significant interaction was reported between the two variables (F(1,199) = 2.19, and p = .141).

To check affirmation manipulation, we asked participants to select three of the 10 sub-categories of

cognitive flexibility that they think humans are superior to machines. It was expected that if an expert suggests that humans are superior to machines in a particular area, then the majority of the participants would report that humans are indeed superior in the area of flexibility that the experts mentioned. The results showed that most of the 95 participants in the cognitive openness condition chose the ability to think flexibly (n = 60), the ability to think creatively (n = 54) and the ability to think openly (n = 39) as their answers. These results show that the participants were influenced by the opinions of the experts in concluding their personal responses.

Main Results: Moderated Mediation Analysis

Table 3 presents descriptive statistics for this study. To test our predictions about the interactive effects of threat manipulation, identity threat, and affirmation manipulation on performance, we used PROCESS SPSS macro (Hayes, 2013; Model 14) for moderated mediation. Bootstrap resamples of 5000 were collected to generate a bias-corrected 95% confidence interval (CI) for each indirect effect (Preacher & Hayes, 2004). In our model, the independent variable was threat manipulation (High vs. Low) and the dependent variable was the number of correct answers in the rationality task, with the degree of perceiving identity threat on inferiority on rational thinking as the mediator. Affirmation manipulation was entered as a moderator between the mediator and the dependent variable.

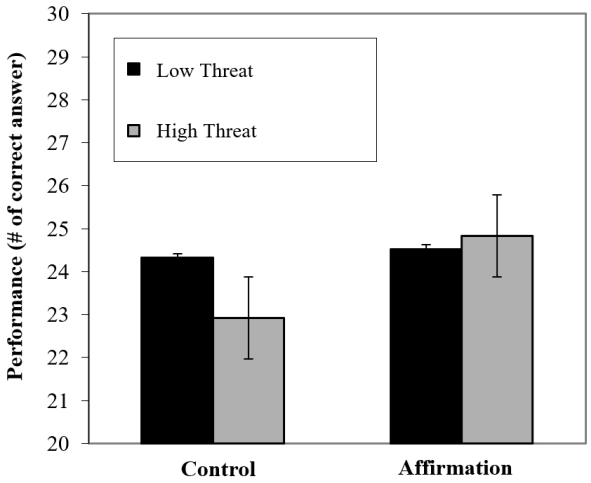


(Figure 1) Moderated mediation model testing the hypothesized mechanism

As illustrated in Figure 1, there was a significant effect of threat manipulation on identity threat (B = 1.56, SE = .19, 95% CI = [1.1839, 1.9301], t(201) = 8.23, and p < .0001) with participants in the high threat condition had significantly higher level of identity threat (M = 4.59,

and SD = 1.35) than those in the low threat condition (M = 3.03, and SD = 1.35). Additionally, the model revealed that the relationship between identity threat and the performance was moderated by the affirmation manipulation (B = .56, SE = .27, 95% CI = [.0161, 1.0890], t(201) = 2.03, and p = .04). For participants who were not treated with affirmation, performance was statistically mediated by identity threat on rationality performance (B = -.81, SE = .41, 95%, and CI = [-1.7410, -.0774]). Yet, no such relationship was found for participants under the affirmation condition (B = .06, SE = .26, 95%, and CI = [-.4581, .5658]).

As a next step, we used a multiple regression analysis to probe the interaction between identity threat and affirmation found in our moderated-mediation model. The two variables were included as predictors in the first step and the interaction term of these variables was entered during the second step. Results revealed that when participants were affirmed, there were no identity threat-based differences in their performances (B = .08, SE = .18, 95% CI = [-.2778, .4432], t(199) = .45, and p = .65). Among participants without affirmation, however, identity threat significantly decreased performance on rational task (B = -.47, SE = .20, 95% CI = [-.8615, -.0698], t(199) = -2.32, and p = .02) (see Figure 2).



(Figure 2) Total number of correct answer that participants solved as a function of identity threat level and affirmation manipulation. For identity threat, high threat refers to values 1 standard deviation above the mean, and low threat refers to values 1 standard deviation below the mean (For affirmation, dummy coded; 1=affirmation, 0=control). Error bars represent ±1 SEM

Discussion

Past research found that a group competition can elicit identity threat and as a line of corollary consequence, threat-managing strategies and performance loss (Becker, 2012; Crocker et al., 1998; Derks, van Laar, & Ellemers, 2007; Osborne, 1997; Schmader, Major, Eccleston, & McCoy, 2001). We examined these basic predictions in a yet undiscovered intergroup context, the competition between humans and machines. Results suggest that identity threats from machines lead people to show performance loss in threat-related tasks. Consistent with past research, these processes were not found when the threat was diminished by affirmations.

Theoretical Implications

The current study holds the potential to contribute to the academia by broadening the scope of applicability of the principles of SIT. While lines of research attempting to apply SIT to human-machine relations are only recently burgeoning (e.g., Fraune, Nishiwaki, Šabanović, Smith, & Okada, 2017; Fraune, Šabanović, & Smith, 2017; Deligianis, Stanton, McGarty, & Stevens, 2017), this study is distinct as it deals with the performance outcome of identity threats and the effect of group-affirmation as a threat-managing strategy. By explaining the impact of intellectual machine as a collective threat, a range of identity-managing mechanisms can be applied to the human-group comparison phenomenon.

Furthermore, what does it mean to identify with the human species as a whole? The human-level identity may be relatively new compared to individual and group-level identities, but in fact, we can find it easily from our psychological reality. For example, human-level identity has often risen from comparisons with animals. Most of the time, humans have a higher status than animals, do not experience threat since we consider ourselves superior to them. Competition with a machine in the Go-match would probably be one of the rare occasions where humans as a group are deemed to be the inferior counterpart in the real world. However, such experience will continue to increase in the future. On the other hand, given that some parts of the human body have blurred boundaries with mechanical and artificial intelligence, the question of which part of the human identity would remain to be unique in the future will become much more prominent. This study provides a starting point for how coping with human identity threats occurs.

Future Directions

Alternative Domains and Forms of Group-affirmation

Future studies may address of which domain of human attributes would show the strongest affirmative effect in buffering threats. For instance, attributes that are less relevant to cognitive ability, such as morality or emotional sensitivity, may serve the most affirmative effect under rationality threat. Relatedly, previous study has shown the effective of affirmation in morality (e.g., playing clean game) in relieving group-level threats (Lalonde, 1992).

Further, as shown in the slogan "Black is beautiful," group members can place a new value on their alternative attributes (Tajfel, 1981). For instance, when they face threats in rationality domain, people may embrace the value of errors in human mind. Future studies should observe other expected forms of group affirmation against threats from the machines.

Emotion-brain Response Study

The match between AlphaGo and Sedol-Lee can also stimulate multi-disciplinary research between psychology and neuroscience.

According to previous research, the level of brain response from certain stimulus differs from individual to individual and can be predicted by the trait of an individual. For instance, different individuals have different levels of sensitivity to threats, which can be predicted from brain responses when given a threatening stimulus (Cools et al., 2005). Thus, we can attempt to connect individual's brain response to perceived threats to human identity elicited from AI and contrast brain responses to emotional stimulus between individuals in future studies.

Specifically, individual differences in emotional response to the Go match can be obtained from the results of previous studies. In previous research (Bae et al., 2017), two emotional measures frequently used in psychology-positive and negative scales (Watson et al., 1998; Lee Hyun-hee et al., 2003) and mood scales (McNair et al., 1992; Kim Eui-Joong et al., 2003) were used to measure the emotional state of study participants before and after the Go match. By substituting the measured data into Russel and his colleague's Circumplex Model (Yick et al., 2011), we can quantify the emotional state into azimuthal values defined in two dimensions (arousal, valence) of a highly quantified Core Affect (CA) space. The dimensions of the same Circumplex Model can then be studied by brain imaging techniques through the Multivariate Pattern Classification (Huth et al., 2012; Ryu et al., In

preparation). Thus, the emotional adjectives felt by individuals after each Go match or the intensity of those feelings can be quantified to the azimuth value of the Circumplex Model or its azimuth intensity, and this positive emotional state can be measured once again through observing patterns of brain activity. This is expected to provide a very important link among data from sociological, psychological, and neuroscientific computational interpretations of the emergence of artificial intelligence.

Furthermore, the measurements of brain activity and individual differences in emotional response to the Go match can be used to identify neural substrates of such personal traits. Previous neuroimaging studies have been identified various neural substrates of human personal traits (Krastev et al., 2016). The measurements and correlation analysis may identify both functional and structural neural substrates of the brain. This finding may provide a potential neuroscientific explanation for such traits.

Social Network Study

Group identity, threatened or re-focused, arguably played a pivotal role in the performance of rationality task in our experiment. One important source of group identity is a person's everyday interactions with significant others, namely, the quality of social relations that help shape his or her understanding of oneself and the social environment. Persons situated in cohesive social relations may develop a strong sense of belonging to their group relative to those embedded in non-cohesive relations such as weak ties. Accordingly, perceived identity threat should vary with the type of social relations that a person maintains. One way to uncover the relational variation in the performance-loss hypothesis, we obtained relational data for a sub-sample (N = 81) and examined whether threat-induced performance was associated with the type of a person's social relations, which was measured by ego network density in her task-advice network (M = .35). A low value of ego network density, typically lower than one third, indicates the non-cohesive social relations that connect individuals of dissimilar attitudes and task-related information (Burt, 1992; Lin, 2001).

The mean for task-performance in this subsample was 24.59, which was largely equivalent to the sample mean. For persons having cohesive relations in the non-affirmation condition (N = 31), the main effect, which is a threat-induced performance loss, was observed at a 0.05 significance level (p = 0.02) and yet in the opposite direction. For the affirmation condition (N = 50), not a notable pattern was identified (See, Figure 3).

	Ego Network Density	
	High	Low
Threat Manipulation		
High Threat	25.28(14)	25.4(5)
Low Threat	23.92(7)	22.4(5)

(Figure 3) Relational Variation in the Non Affirmation Condition

Note: Number of observations in the parenthesis

Apparently, individuals exerted more efforts in the experimental task when their group identity was threatened and exhibited better performance on the task, which was being viewed as the best application of so-called smart machines. One possible implication would be that identity threat may motivate those who belong to a cohesive group to engage actively in the task domain previously related to the superiority of their group.

Practical Implications

Socio-political

Intellectual games have been traditionally used to clarify boundaries between human and non-human beings. Especially, go has had a reputation as one of the most complicated games and thus has been regarded as the last game for humans to be caught up by machine intelligence (Silver et al., 2016). However, human superiority seems invaded in intellectual matches on go against rapidly advancing machine intelligence. For instance, 60 top professionals including the world champion had tried to regain human superiority on go but failed to win even in one game.

Threats from the human-level machines would not be limited to the rational thinking dimensions. Right after the Go match, Dr. Silver who hosted the match said he began developing an AI system that provides customized medical services by generalizing AlphaGo's learning algorithms. He also revealed that AlphaGo is currently collaborating with the National Institutes of Health to provide personalized health care by learning individual medical data (Kim Ji-min, 2016).

Following the victory of Alphago in the Go game, predictions arose that AI will take on the role of human beings and create social chaos. The role of human labor in the manufacturing industry has long been replaced by specialized machines, and this trend is expected to rapidly spread across other

industries with the development of AI technology.

Thus, it would be beneficial to prepare for the potential social shocks. If the confrontation between humans and machines captured in the go match represents the expected antagonism between humans and machines, we will be able to identify individual differences and social phenomena based on the results of this study and devise social policies in preparation. Also, increasing number of studies on the interaction between humans and AI will enable us to further anticipate differences in people's perceptions and attitude and design the best policy for them.

Economy Industrial

In 2015, AI (Artificial Intelligence) was selected among future core technologies from 15 fields of excellent industrial value presented by the Future Preparatory Committee under the Ministry of Science, ICT and Future. Its value is expected to grow exponentially in the future due to its immeasurable potential applications to other fields and technologies such as big data and IoT (Lee Kwang-hyung et al.. 2015). According to the report, AI is particularly closely associated with economic and industrial values, such as manufacturing revolution, polarization of industrial structure, low growth and growth strategy. The application of AI is exploding in the industrial sector, and this trend is expected to continue.

As AI technology penetrates into various industries, the number of companies launching AI-enabled products and related industries are expected to increase. Companies entering new markets must perform marketing analysis to anticipate consumer reactions to their products with AI technology and project market responses to their products. We expect this study concerning individual's perception and attitude toward artificial intelligence can be used as a stepping stone to provide invaluable insight to industry.

Human-Machine Interaction

What would interactions with future machines and AI be like and what strategies can be implemented? The number of qualities that machines possess superiority to humans can be greater in the near future. Threats arising from superiority of these machines may lead to a decline in human motivation and performance, especially in the threatened dimensions. When machines that are more sociable than humans are developed, will we dismiss our need for social interactions? In films that have already extrapolated the development of technology, we have been warned of weakening social

relations (“Her”) and the feeling of inferiority (“Gattaca”) in the form of future humanity.

Furthermore, if the day comes when machines dominate all aspects of humanity, that is, when we have nothing left to be affirmed, will we eventually lose all motivation? Inferring from the results of this study makes such a conclusion possible: humans now seek for elsewhere to be affirmed, but it is probable that nothing will be affirmable in the future. This study suggests a social psychological basis for appropriately regulating machines and the relevant industry so that elsewhere will no longer be non-existent.

Our study also suggests a direction for future A.I design, which targets to develop a more human like A.I. The development for such A.I systems requires inputs that are identified as superiority of humans over A.I systems. These inputs may also suggest the directions for future A.I development. One such direction is developing emotional A.I (Lee, & Gurnkl, & Düsentrieb, 2015; Picard, 2004). Our research strategy may be helpful in identifying core properties of human emotion that are required for the developing of future emotional A.I.

Conclusion

The Google Deep Mind Challenge Match was an event that illustrates the intergroup competition between human and machine. To analyze, understand, explain, and predict how the dynamic interplay of human-machine relation will affect human psychology and society in general is a complicated task. The current study provided a stepping stone for preliminary explanations of the performance-loss as a corollary consequence of machine threat and the effect of group affirmation as a threat-buffer. Building on the current findings, research observing the human-machine interaction is called upon. For example, in the case of educational scenes that are continuously increasing interaction between human and machines, people’s positive attitudes and trust in Machines are very important. If we can predict what psychological consequences such interactions will bring, we can provide a very powerful communicative solution. In order to derive deeper implications, an interdisciplinary approach that combines brain science, and business analysis is required.

References

Bae, J., Cha, Y.-J., Lee, H., Lee, B., Baek, S., Choi, S., & Jang, D. (2017). Social networks and inference about unknown events: A case of the match between Google's AlphaGo and Sedol Lee. *PloS One*, *12*(2), e0171472.

Becker, J. C. (2012). The system-stabilizing role of identity management strategies: Social creativity can undermine collective action for social change. *Journal of Personality and Social Psychology*, *103*(4), 647-662.

Branscombe, N. R., Ellemers, N., Spears, R., & Doosje, B. (1999). The context and content of social identity threat. In N. Ellemers, R. Spears, & B. Doosje (Eds.), *Social identity: Context, commitment, content* (pp. 35-58). Oxford: Blackwell Science.

British Science Association. (n.d.). *One in three believe that the rise of artificial intelligence is a threat to humanity*. Retrieved from https://www.britishscienceassociation.org/news/rise-of-artificial-intelligence-is-a-threat-to-humanity

Burt, R. S. (2009). Structural holes: The social structure of competition. Harvard university press.

Crocker, J., Major, B., & Steele, C. (1998). Social stigma: The psychology of marked relationships. In *The handbook of social psychology* (Vol. 2, pp. 504-553). Boston: McGraw-Hill.

Deligianis, C., Stanton, C. J., McGarty, C., & Stevens, C. J. (2017). The impact of intergroup bias on trust and approach behaviour towards a humanoid robot. *Journal of Human-Robot Interaction*, *6*(3), 4-20.

Derks, B., Scheepers, D., Van Laar, C., & Ellemers, N. (2011). The threat vs. challenge of car parking for women: How self-and group affirmation affect cardiovascular responses. *Journal of Experimental Social Psychology, 47*(1), 178-183.Derks, B., van Laar, C., & Ellemers, N. (2006). Striving for success in outgroup settings: Effects of contextually emphasizing ingroup dimensions on stigmatized group members' social identity and performance styles. *Personality and Social Psychology Bulletin*, *32*(5), 576-588.

Eyssel, F., & Kuchenbrandt, D. (2012). Social categorization of social robots: Anthropomorphism as a function of robot group membership. *British Journal of Social Psychology*, *51*(4), 724-731.

Ferrari, F., Paladino, M. P., & Jetten, J. (2016). Blurring human-machine distinctions: Anthropomorphic appearance in social robots as a threat to human distinctiveness. *International Journal of Social Robotics*, *8*(2), 287-302.

Fraune, M. R., Nishiwaki, Y., Sabanović, S., Smith, E. R., & Okada, M. (2017). Threatening flocks and mindful snowflakes: How group entitativity affects perceptions of robots. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 205-213). New York: ACM.

Fraune, M. R., Šabanović, S., & Smith, E. R. (2017). Teammates first: Favoring ingroup robots over outgroup humans. In *26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), 2017* (pp. 1432-1437). Lisbon: IEEE.

Gunn, G. R., & Wilson, A. E. (2011). Acknowledging the skeletons in our closet: The effect of group affirmation on collective guilt, collective shame, and reparatory attitudes. *Personality and Social Psychology Bulletin, 37*(11), 1474-1487.

Hogg, M. A., Abrams, D., Otten, S., & Hinkle, S. (2004). The social identity perspective: Intergroup relations, self-conception, and small groups. *Small Group Research*, *35*(3), 246-276.

Kim Eui-Joong et al. (2003). han-kwuk-phan ki-pwun-sang-thay-chek-to(K-POMS) uy phyo-cwun-hwa-wa sin-loy-to-wa tha-tang-to phyeng-ka 한국판 기분상태척도(K-POMS)의 표준화와 신뢰도와 타당도 평가 [Standardization, Reliability, and Validity Evaluation of the Korean Version of Mood Scale(K-POMS)]. *swu-myen-ps ceng-sin-sayng-li, 10*(1), 39-51.

Kim Ji-min. (2016. 3. 8). al-pha-ko kay-pal-ca “kin ye-ceng-uy ches kel-um-ss-uy-lyo AI kot na-on-ta” 알파고 개발자 “긴 여정의 첫 걸음…의료 AI 곧 나온다” [AlphaGo Developer “First Steps on Long Journey… Medical AI Coming Soon”]. Moneytoday. URL: https://news.mt.co.kr/mtview.php?no=2016030812325896447&VN.

Krastev, S., McGuire, J. T., McNeney, D., Kable, J. W., Stolle, D., Gidengil, E., & Fellows, L. K. (2016). Do political and economic choices rely on common neural substrates? A systematic review of the emerging neuropolitics literature. *Frontiers in psychology, 7*, 264.

Lalonde, R. N. (1992). The dynamics of group differentiation in the face of defeat. *Personality and Social Psychology Bulletin*, *18*(3), 336-342.

Lee Hyun-hee et al. (2003). han-kwuk-phan ceng-cek ceng-se mich pwu-cek ceng-se chek-to(Positive Affect and Negative Affect Schedule; PANAS) uy tha-tang-hwa yen-kwu 한국판 정적 정서 및 부적 정서 척도(Positive Affect and Negative Affect Schedule; PANAS)의 타당화 연구[Validation of the Korean Version of Positive Affect and Negative Affect Schedule]. *Korean Journal of Clinical Psychology, 22*(4), 935-946.

Lee, P & Gurnkl, D & Düsentrieb, Daniela. (2015). Why Truly Intelligent Machines Need Emotions. *International Journal of Artificial Intelligence. 3*.

Lee Kwang-hyung et al. (2015). mi-lay-i-syu pwun-sek po-ko-se 미래이슈 분석 보고서 [Future Issue Analysis Report]. Ministry of Science, ICT and Future Planning.

Lin, N. (2002). Social capital: A theory of social structure and action (Vol. 19). Cambridge university press.

Picard, R. W. (2004). Toward Machines With Emotional Intelligence. In *The Science of Emotional Intelligence:*

Knowns and Unknowns. 29-30. 10.1093/acprof:oso/9780195181890.003.0016.

Sherman, D. K., Kinias, Z., Major, B., Kim, H. S., & Prenovost, M. (2007). The group as a resource: Reducing biased attributions for group success and failure via group affirmation. *Personality and Social Psychology Bulletin*, *33*(8), 1100-1112.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., … Lanctot, M. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*(7587), 484-489.

Steele, C. M. (1988). The psychology of self-affirmation: Sustaining the integrity of the self. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 21, pp. 261-302). San Diego, CA: Academic Press.

Tajfel, H. (1981). *Human groups and social categories: Studies in social psychology*. CUP Archive.

Tajfel, H., & Turner, J. C. (1979). An integrative theory of intergroup conflict. In W. G. Austin, & S. Worchel (Eds.), *The social psychology of intergroup relations* (pp. 33-37). Monterey, CA: Brooks/Cole.

Tajfel, H., & Turner, J. C. (1986). The social identity theory of intergroup behavior. In S. Worchel & W. Austin (Eds.), *Psychology of intergroup relations* (pp. 7-24). Chicago: Nelson-Hall.

Tajfel, H., & Wilkes, A. L. (1963). Classification and quantitative judgement. *British journal of psychology, 54*(2), 101-114.

Van Laar, C., Derks, B., & Ellemers, N. (2013). Motivation for education and work in young Muslim women: The importance of value for ingroup domains. *Basic and Applied Social Psychology*, *35*(1), 64-74.

Yeo, I. (2017). *Human perception on artificial intelligence: Blessing or threat?* (Master's thesis). Seoul National University.

Yogeeswaran, K., Złotowski, J., Livingstone, M., Bartneck, C., Sumioka, H., & Ishiguro, H. (2016). The interactive effects of robot anthropomorphism and robot ability on perceived threat and support for robotics research. *Journal of Human-Robot Interaction*, *5*(2), 29-47.

1차 원고 접수: 2019. 09. 06
1차 심사 완료: 2019. 09. 27
2차 원고 접수: 2019. 10. 02
2차 심사 완료: 2019. 10. 04
3차 원고 접수: 2019. 10. 04
최종 게재 확정: 2019. 10. 04

(요 약)

기계의 정체성 위협에 대항하기: 집단 가치 확인이 인간의 지적 수행 저하에 미치는 효과

차 영 재[1)] 백 소 정[1)] 이 형 석[2)] 배 종 훈[3)]
이 종 호[4)] 이 상 훈[5)] 김 건 희[6)] 장 대 익[7)†]

[1)]인지과학 협동과정 [2)]과학사 및 과학철학 협동과정 [3)]경영전문대학원 [4)]전기·정보공학부
[5)]뇌인지과학과 [6)]컴퓨터공학부 [7)]자유전공학부
* 모든 저자는 서울대학교에 소속됨

인공 지능으로 인한 정체성 위협은 지능 과제에 대한 동기 및 수행을 저해할 수 있다. 본 연구는 실험 기법을 활용하여 개인의 지능 과제 수행 저하 현상이 인공 지능으로 인한 위협에 노출됨으로써 나타나는지 조사하였다. 또한 본 연구는 집단 정체성 확인(group identity affirmation)이 과제 수행 저하 현상을 완화해줄 수 있는지 확인하였다. 구체적으로, 인공지능 위협이 높은 조건에서는 낮은 조건에서보다 지적 과제 수행이 낮을 것으로 예측하였다. 또한 이와 같은 수행 저하 효과는 집단 확인 조건에서 나타나지 않을 것으로 예측하였다. 대학생 참가자 210명을 대상으로 실험 연구를 시행하여 예상과 일관된 결과를 발견하였다. 인공지능으로 인한 정체성 위협은 참가자의 지적 과제 수행을 떨어뜨렸으며, 이와 같은 수행 저하 현상은 집단 가치 비 확인 조건에서 발견됐지만 집단 가치 확인 조건에서는 발견되지 않았다. 논의에서는 이론적·실용적 함의를 다루었다.

주제어 : 사회정체성, 정체성 위협, 수행 손실, 인간-기계 경쟁, 집단 가치 확인, 집단 간 관계