Emotion-Based Agent-based Modeling and Simulation of Hybrid Cyber Attack Scenarios

Jeongkeun Shin[1], Ellen Kim[2], Clarissa Chau[2], L. Richard Carley[1], and Kathleen M. Carley[1]

[1] CASOS Center, Carnegie Mellon University, Pittsburgh PA 15213, USA
[2] Troy High School, Fullerton CA 92831, USA
{jeongkes,lrc,carley}@andrew.cmu.edu
{ellenkim236,cchau0908}@gmail.com

Abstract. **Keywords:** Agent-based Modeling and Simulation · Emotion · Human Factors · Cybersecurity · MITRE ATT&CK.

1 Introduction

Organizations defend against cyberattacks through a combination of technical controls and human judgment. An intrusion detection system may generate an alert, but a person ultimately decides whether to act on it. Security awareness training teaches employees to recognize phishing, yet they must recall that training when a suspicious email arrives. The effectiveness of a defense therefore depends not only on what has been deployed but also on the condition of the people operating it.

Emotion is among the most immediate determinants of that condition. Organizational research consistently shows that affective state shapes job performance [2,25], and there is little reason to expect security work to be an exception [11]. Emotion also fluctuates over hours and days, far more rapidly than the organizational attributes through which cybersecurity models typically represent people. It may therefore explain why the same organization, with the same staff and tools, withstands an attack effectively on one day but poorly on another.

This paper asks two questions. First, how much of an organization's cyberattack damage is attributable to the emotional state of its members relative to the technical skill of its security staff? Second, how do these two factors interact? We address these questions by making emotion the source of job performance in an agent-based model of an organization under attack and evaluating the resulting model through a complete factorial design. The orthogonality of this design allows damage to be attributed to each factor without overlap.

2 Related Work

Cybersecurity simulation recreates cyberattack situations that are difficult to reproduce and impossible to repeat in the real world within an isolated virtual environment, allowing the effectiveness of different conditions and defensive

strategies to be tested experimentally [12]. Early research focused on the technical aspects of the problem, particularly how faithfully network environments and attack techniques could be represented in a simulation. As social engineering became an increasingly prominent route of intrusion, however, human factors also emerged as an important object of simulation.

Blythe et al. modeled the behavior of cyber defenders and simulated how it changes under fatigue and stress [5], while Burns et al. drew on social science theory to propose an approach to modeling behavioral change in human agents [6].

A parallel stream of research has attributed security failures not to a lack of training but to systematic distortions in reasoning. Andrade et al. proposed a conceptual model explaining why even trained users fall for phishing attacks through heuristics such as authority and urgency framing [4]. Mansaray et al. identified biases that distort defenders' security decision making and proposed strategies to mitigate them [14]. Aggarwal et al. showed that classic biases such as recency bias and loss aversion carry over into cyber contexts from the attacker's perspective [1]. While these biases are relatively stable tendencies that shape which actions individuals choose, the emotional states modeled in this study fluctuate from day to day and alter the amount of deliberative effort devoted to carrying out those actions. Emotion therefore serves as one condition under which latent biases manifest in behavior.

To the best of our knowledge, however, no previous study has quantified how the emotional state of organizational members affects an organization's cybersecurity resilience. This paper addresses that gap.

3 Cyber Attack & Defense Model

The attack model represents a ten-day hybrid campaign in which data exfiltration and denial of service occur in parallel. **APT 28** was selected because the group has a documented history of conducting both phishing-based data exfiltration and Denial-of-Service operations, allowing both attack streams to be modeled consistently as the activities of a single threat actor [17].

As illustrated in Figure 1, the simulation operates on the attack flow network constructed by Shin et al. [21], where ATT&CK techniques [26] are represented as nodes and observed transitions between techniques as edges. Given a source and a final objective technique, the attacker advances by repeatedly sampling the next technique from the outgoing edges of its current position. Candidate successors are restricted to those that retain a path to the objective [22].

The two attack streams run concurrently over ten days. The first pursues data exfiltration from a phishing entry point (T1566.001 to T1567), while the second launches a DoS campaign every three hours (T1595.002 to T1498).

The target organization continuously monitors its network traffic using an intrusion detection system designed to detect attacker activity. We assume that the organization's network environment resembles that represented by the NSL-

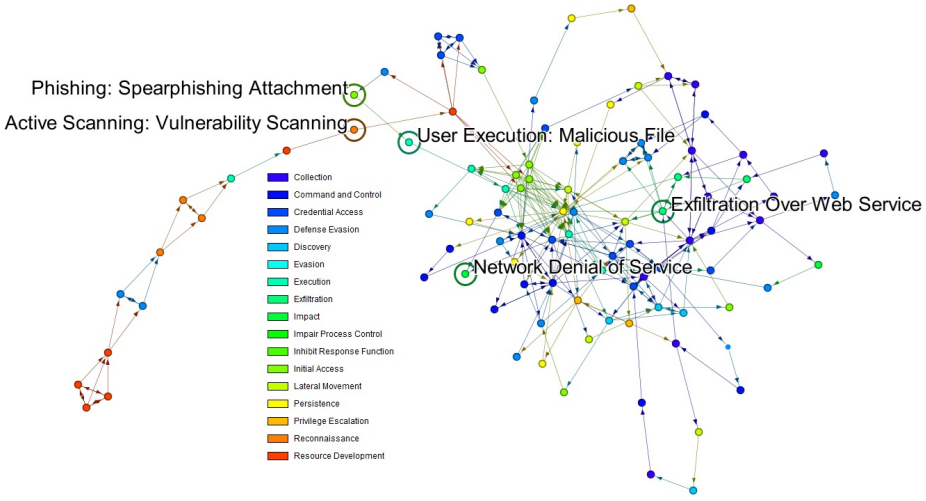


Fig. 1. Network Model of Attack Scenario of APT 28 [17] visualized by ORA [3,7].

KDD dataset [27]. A Bayes Net classifier [9] trained on NSL-KDD [27] serves as the detection engine.

Table 1. Performance of ML-based Intrusion Detection System (IDS), given as the true positive rate (TPR), with Mapping Between MITRE ATT&CK Techniques and NSL-KDD Labels.

Technique	NSL-KDD Labels	Detection Rate (TPR)
T1498	neptune, smurf, udpstorm	0.946
T1595.002	satan, mscan, saint	0.944
T1110.001	guess_passwd, snmpguess	0.845
T1190	imap, phf, sendmail, named	0.400
T1083	spy	0.000
T1105	warezclient	0.994
T1210	worm	0.000
T1068	buffer_overflow, loadmodule, ps, xterm	0.778
T1014	rootkit	0.000

Because NSL-KDD attack labels [27] and MITRE ATT&CK techniques [26] use different taxonomies, we relied on a mapping established in prior work [23]. Using this mapping, we identified the MITRE ATT&CK techniques known to have been used by APT 28 [17] that correspond to an NSL-KDD attack label [27]. Table 1 reports the rate at which the trained model detects each corresponding label. Of the 96 techniques in the APT 28 attack flow network [21], 9 can be mapped in this manner. The execution of any remaining technique therefore

produces no signal for the IDS. When the IDS detects an attack, it immediately alerts the organization's IT security officer. T1595.002 (Active Scanning), however, is excluded from alert generation in our model because it represents reconnaissance conducted outside the network perimeter, where detection does not allow the activity to be stopped.

4 Human & Organization Model

In our simulation model, every human agent in the target organization is influenced by the levels of Plutchik's eight basic emotions [18]. Human agents with the same cybersecurity knowledge may exhibit different levels of job performance depending on their current emotional state, leading to differences in their behavior and proactiveness during cybersecurity situations. Consequently, the damage caused by the same attack can vary substantially. This paper derives each agent's job performance from its Plutchik emotion state [18], models job performance as the driver of employee behavior during an attack, and measures the resulting damage through simulation.

No existing study directly relates all eight of Plutchik's emotions [18] to job performance. We therefore obtained the reported effect of each emotion on job performance from a separate study, rescaled its magnitude to a 1-to-5 scale, and expressed it as the change in job performance associated with a one-point increase on that scale. The resulting coefficients are Joy (+0.310) [15], Trust (+0.589) [29], Fear (-0.304) [31], Surprise (+0.380) [13], Sadness (-0.264) [19], Disgust (-0.279) [30], Anger (-0.074) [28], and Anticipation (+0.695) [16]. Because each coefficient was derived from a different study that considered a different set of variables, these values should be interpreted as a face validation [20] of the direction and approximate magnitude of each effect.

An agent's emotion score is calculated as the weighted sum of its eight emotions, with each emotion offset by the floor of the emotion scale and multiplied by its corresponding coefficient. The resulting sum has an attainable range of [-3.684, 7.896], which is linearly rescaled to the job performance scale [0, 5] to obtain the agent's final job performance. Following the organizational model of Shin et al., the organization consists of 235 end user agents and one IT security officer agent [24]. Each end user agent is assigned an individual phishing susceptibility value before security education (PSBE) and after security education (PSAE), based on values derived by Shin et al. using the phishing susceptibility model of Eftimie et al. [10,24].

All users have completed security training, so PSAE serves as the baseline. However, recall of that training at the moment of an attack is not guaranteed. When an agent encounters a phishing email, higher emotion-based job performance increases the likelihood of recalling the training, in which case PSAE applies. An agent with lower job performance is more likely to forget the training and revert to PSBE [24]. The probability of being deceived therefore depends not simply on whether the user has received training, but on the user's emotional state at the time of the attack and the resulting level of job performance.

In addition to job performance, the IT security officer agent has a second human factor that we call *cyber capability*. This factor represents the officer's defensive skill and ranges from 1 to 5. A higher value increases the likelihood that the officer identifies an ongoing attack and stops it early after examining an IDS alert. It also increases the likelihood of detecting an intrusion on an end user device during an inspection and disconnecting the compromised device. Job performance, by contrast, determines how actively the officer works. Higher job performance increases the probability that an IDS alert is examined rather than ignored and raises the number of end user devices inspected each day. The two factors are independent, allowing a diligent but unskilled officer and a skilled but inattentive officer to produce distinct outcomes.

5 Virtual Experiments

We constructed the virtual experiment as a full factorial design to isolate the effects of emotion and defensive skill on the resulting damage. Table 2 summarizes the experimental conditions. Each of Plutchik's eight emotions [18] takes one of five levels, and the IT security officer's cyber capability also takes one of five levels, yielding $5^8 \times 5 = 1,953,125$ conditions. Within each condition, the same emotion levels are assigned to all 235 end user agents and the IT security officer, whereas cyber capability applies only to the IT security officer.

Because the simulation is stochastic, each condition was replicated five times using different random seeds, producing a total of 9,765,625 simulation runs. Each run spans ten days at one minute per tick (14,400 ticks). The two attack scenarios and the IDS detection rates were held constant across all runs. The dependent variables recorded for each run are the total volume of data exfiltrated through T1567 in GB and the cumulative server downtime caused by T1498.

6 Results

Figure 2 presents the response surfaces of both damage measures across job performance and cyber capability. Both measures decline as the two factors increase, but the magnitude of the reduction depends strongly on their combination. In the lowest job performance band (0.0 to 0.5), increasing cyber capability from 1 to 5 reduces exfiltration from 321 GB to 296 GB, a decrease of only 8 percent. In the highest band (4.5 to 5.0), the same increase reduces exfiltration from 160 GB to 60 GB, a decrease of 63 percent. Server downtime exhibits a similar pattern, declining by 5 percent and 49 percent, respectively. The same improvement in defensive capability therefore produces nearly an eightfold difference in its effect on exfiltration depending on the organization's emotional state, indicating that the two factors interact rather than operate additively. Viewed from the opposite direction, even at the lowest level of cyber capability, increasing job performance from the lowest band to the highest halves exfiltration from 321 GB to 160 GB. The human factor therefore functions as an independent defensive resource rather than merely as a multiplier of technical capability.

Table 2. Virtual Experiment Summary.

Independent Variable		
Variable	**Range**	**Implication**
Plutchik Emotion	1–5	Determines agents' job performance based on the levels of the eight emotions and their respective weights.
Cyber Capability	1–5	Represents the ability of the IT security officer agent to accurately detect attacks during an inspection.
Mediating Variable		
Variable	**Range**	**Implication**
Job Performance	0–5	Derived from the eight emotions, it affects end users' recall of cybersecurity training, the probability that the IT security officer inspects IDS alerts rather than ignoring them, and the PC inspection quota.
Dependent Variable		
Variable	**Unit**	**Implication**
Data Exfiltration Damage	GB	Volume of data exfiltrated from compromised PCs through spearphishing campaigns.
Server Downtime Damage	Minute	Cumulative server downtime caused by Denial-of-Service (DoS) attacks.
Control		
Variable	**Specification**	**Implication**
Attacker	–	Conducts spearphishing and DoS attacks.
IDS	9 Techniques	A defensive system that monitors nine attack techniques, five of which generate alerts in practice.
IT Security Agent	1 agent	Triages IDS alerts and proactively inspects individual PCs. When an attack is detected, the affected device is isolated or the DoS attack is stopped.
End User Agent	235	Each agent has a unique phishing susceptibility and a computing device containing 10 GB of data.
Attack Duration	10 days	14,400 ticks.

Table 3 reports regressions of each damage measure on the eight emotions and cyber capability. Because the experiment uses a complete factorial design, every predictor is orthogonal to all others, and the squared standardized coefficients sum exactly to the variance explained by the model. This property allows the contributions of the two types of factors to be separated cleanly. For data exfiltration, cyber capability accounts for 50.0 percent of the explained variance, while the eight emotions collectively account for the remaining 50.0 percent, indicating a balance between the human and technical factors. For server downtime, cyber capability accounts for 74.9 percent and the emotions for 25.1 percent, making the technical factor approximately three times as influential.

This difference reflects the defensive pathways on which each damage measure depends. In defending against exfiltration, emotion-derived job performance operates through three channels, namely recall of training content, inspection of

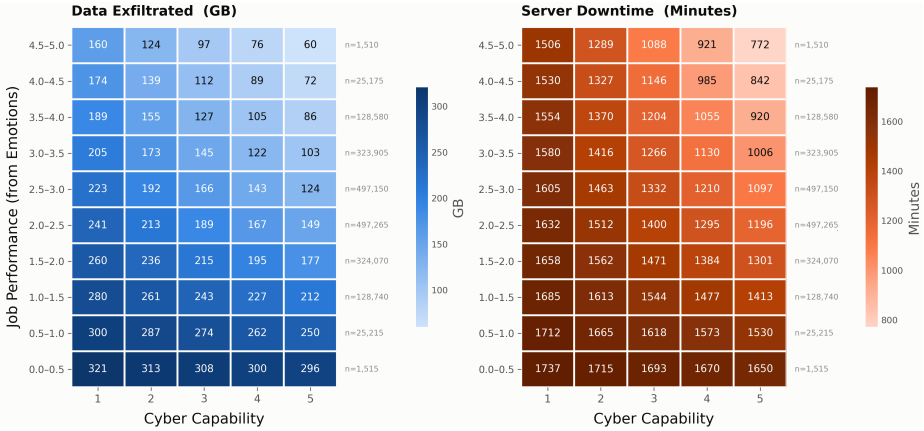


Fig. 2. Damage over job performance (0–5) and cyber capability (1–5): exfiltrated data (GB) and server downtime (min).

Table 3. Regression Coefficients for Data Exfiltrated and Server Downtime

Variable	**Data Exfiltrated (GB)**			**Server Downtime (Min)**		
	B	SE	β	B	SE	β
Intercept	315.8783	0.0774	–	1911.4501	0.1510	–
Cyber Capability	-23.1165	0.0085	-0.549	-116.4550	0.0166	-0.808
Anticipation	-13.9837	0.0085	-0.332	-40.7164	0.0166	-0.283
Trust	-11.8497	0.0085	-0.282	-34.5215	0.0166	-0.240
Surprise	-7.6664	0.0085	-0.182	-22.2831	0.0166	-0.155
Joy	-6.2379	0.0085	-0.148	-18.1706	0.0166	-0.126
Fear	6.1216	0.0085	0.145	17.8263	0.0166	0.124
Disgust	5.6280	0.0085	0.134	16.3186	0.0166	0.113
Sadness	5.3230	0.0085	0.126	15.4925	0.0166	0.108
Anger	1.5066	0.0085	0.036	4.3559	0.0166	0.030

alerts, and inspection of devices. Device inspection does not contribute to server defense, leaving only two channels. The individual emotion coefficients reduce damage in the order of Anticipation, Trust, Surprise, and Joy, while they increase damage in the order of Fear, Disgust, Sadness, and Anger. This ordering and the relative magnitudes reproduce the emotion weights used in the model. The simulation further provides the absolute scale of these coefficients by estimating how much damage, measured in gigabytes or minutes, corresponds to a one-point change in each emotion. For example, a one-point increase in Anticipation reduces exfiltration by 13.98 GB and server downtime by 40.7 minutes.

7 Discussion and Conclusion

This study modeled job performance derived from Plutchik's eight basic emotions [18] as a driver of an organization's cybersecurity response behavior and quantified the influence of emotion on cyberattack damage through 9,765,625 simulation runs. Emotion has explanatory power comparable to that of technical defensive capability, and the two factors interact in shaping the resulting damage. When formulating a cybersecurity defense strategy, the emotional state of organizational members should therefore be treated not as an incidental consideration but as a factor to be weighed alongside technical controls.

In this study, emotional state is assigned uniformly across the organization and held constant throughout the simulation. This deliberate simplification isolates the effect of emotion within a factorial design, but emotional states in real organizations vary across individuals and change in response to events. The study is also limited to a single threat group and two attack scenarios, and the model was not validated against real incident data. Finally, the emotion weights were drawn from separate studies and rescaled to a common scale, so they should be interpreted as a face validation [20] of the direction and approximate magnitude of each effect.

We plan to connect the model to the field of social cybersecurity [8] and examine scenarios in which social cyberattacks, such as disinformation campaigns or opinion manipulation, alter the emotional state of organizational members and thereby degrade the organization's ability to respond to a cyberattack. This approach would support composite scenarios in which social and technical attacks unfold together over time. Additional directions for future work include introducing emotional heterogeneity among members, extending the model to other threat groups, and validating it against real incident data.

Acknowledgements The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported in part by the Minerva Research Initiative under Grant #N00014-21-1-4012 and by the Center for Computational Analysis of Social and Organizational Systems (CASOS) at Carnegie Mellon University. The views and conclusions are those of the authors and should not be interpreted as representing

the official policies, either expressed or implied, of the Office of Naval Research or the US Government.

References

1. Aggarwal, A., Ferreira, M.J., Aggarwal, P., Rajivan, P., Gonzalez, C.: Cognitive biases in cyber attacker decision making: Translating behavioral insights into cybersecurity. In: 2025 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW). pp. 630–645. IEEE (2025)
2. Aitken, J.A., Baines, J.I., Wonders, M.E., Kaplan, S.A., Clark, J.E.: A meta-analytic review of the within-person relationship between affect and job performance. Journal of Organizational Behavior **47**(4), 685–709 (2026)
3. Altman, N., Carley, K.M., Reminga, J.: ORA user's guide 2020. Technical Report 2, Carnegie Mellon University, Institute for Software Research International, Pittsburgh, PA (2020)
4. Andrade, R., Cazares, M.F., Cordova-Pintado, D.: The human factor in cybersecurity: A conceptual model integrating cognitive biases. In: World Congress in Computer Science, Computer Engineering & Applied Computing. pp. 3–13. Springer (2025)
5. Blythe, J., Botello, A., Sutton, J., Mazzocco, D., Lin, J., Spraragen, M., Zyda, M.: Testing cyber security with simulated humans. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 25, pp. 1622–1627 (2011)
6. Burns, A., Posey, C., Courtney, J.F., Roberts, T.L., Nanayakkara, P.: Organizational information security as a complex adaptive system: insights from three agent-based models. Information Systems Frontiers **19**(3), 509–524 (2017)
7. Carley, K.M.: ORA: A toolkit for dynamic network analysis and visualization (2014)
8. Carley, K.M.: Social cybersecurity: an emerging science. Computational and mathematical organization theory **26**(4), 365–381 (2020)
9. Cooper, G.F., Herskovits, E.: A bayesian method for the induction of probabilistic networks from data. Machine learning **9**(4), 309–347 (1992)
10. Eftimie, S., Moinescu, R., Răcuciu, C.: Spear-phishing susceptibility stemming from personality traits. IEEe Access **10**, 73548–73561 (2022)
11. Helkala, K., Knox, B., Jøsok, Ø., Knox, S., Lund, M.: Factors to affect improvement in cyber officer performance. Information & Computer Security **24**(2), 152–163 (2016)
12. Kavak, H., Padilla, J.J., Vernon-Bido, D., Diallo, S.Y., Gore, R., Shetty, S.: Simulation for cybersecurity: state of the art and future directions. Journal of Cybersecurity **7**(1), tyab005 (2021)
13. Kerkhoven, O.: The effect of surprise on creative performance and surprise bias. Master's Thesis, Tilburg University (2021)
14. Mansaray, E., Abdulnabi, M., Nathan, Y., Alizadeh, S., Ali, A., AL-Maari, A.A., Ghanem, W.A.H.: Understanding cognitive biases in security decision-making and proposing effective mitigation strategies. In: International Conference on Microelectronics, Electromagnetics and Telecommunication. pp. 611–624. Springer (2024)
15. Met, I., Erkoç, A., Ipek, B., Kızıldere, N.A.: Being a game changer with happy workers: The ziraat bank example. Heliyon **9**(7) (2023)
16. Miralam, M., Ali, N.: Employees self-efficacy, optimism, hope, resilience and demographic variables as the determinants of job performance. International Journal on Emerging Technologies **11**(3), 336–343 (2020)

17. MITRE ATT&CK: APT28, G0007. https://attack.mitre.org/versions/v17/groups/G0007/ (2025), version 17; accessed 2026-08-11
18. Plutchik, R.: A general psychoevolutionary theory of emotion. In: Theories of emotion, pp. 3–33. Elsevier (1980)
19. Rispens, S., Demerouti, E.: Conflict at work, negative emotions, and performance: A diary study. Negotiation and Conflict Management Research **9**(2), 103–119 (2016)
20. Sargent, R.G.: Verification and validation of simulation models. In: Proceedings of the 2009 Winter Simulation Conference (WSC). pp. 162–176. IEEE (2009)
21. Shin, J., Zhai, S., Carley, L.R., Carley, K.M.: Attack flow network models of mitre att&ck groups and campaigns. Technical Report CASOS Technical Report 2025, CASOS (2025), forthcoming
22. Shin, J., Carley, L.R., Carley, K.M.: Cyber attack flow intelligence network: Backbone analysis and defense strategy optimization. IEEE Transactions on Network Science and Engineering (2026)
23. Shin, J., Carley, R., Carley, K.: Simulation-based evaluation of ids monitoring strategies for denial-of-service defense using attack flow analysis. In: Proceedings of the 2026 Winter Simulation Conference (Dec 2026), forthcoming
24. Shin, J., Zhai, S., Carley, L.R., Carley, K.M.: Simulating cyber defense: the impact of phishing training and system updates on mitigating damage from hybrid phishing and watering hole attacks. The Journal of Defense Modeling and Simulation p. 15485129251365259 (2025)
25. Shockley, K.M., Ispas, D., Rossi, M.E., Levine, E.L.: A meta-analytic investigation of the relationship between state affect, discrete emotions, and job performance. Human Performance **25**(5), 377–411 (2012)
26. Strom, B.E., Applebaum, A., Miller, D.P., Nickels, K.C., Pennington, A.G., Thomas, C.B.: Mitre att&ck: Design and philosophy. Technical report, The MITRE Corporation (2018)
27. Tavallaee, M., Bagheri, E., Lu, W., Ghorbani, A.A.: A detailed analysis of the kdd cup 99 data set. In: 2009 IEEE symposium on computational intelligence for security and defense applications. pp. 1–6. Ieee (2009)
28. Umbra, R., Fasbender, U.: The daily relations between workplace anger, coping strategies, work outcomes, and workplace affiliation. Frontiers in Psychology **16**, 1538914 (2025)
29. Varshney, D., Varshney, N.K.: Measuring the impact of trust on job performance and self-efficacy in a project: evidence from saudi arabia. Journal of Applied Business Research **33**(5) (2017)
30. Yang, R., Li, N., Fan, Y., Yuan, Z., Lin, C., Shen, Y., Yang, L.: The relationship between clinical nurses' feelings of underqualification and work withdrawal behavior: the multiple mediating roles of self-disgust and self-efficacy. Frontiers in Psychology **16**, 1588917 (2025)
31. YEŞİL, S., Mavi, Y.: Covid-19 korkusu'nun duygusal emek, çalışan performansı, iş tatmini, yaşam doyumu üzerine etkisi: Bir alan araştırması. Journal of Yaşar University/Yaşar Üniversitesi E-Dergisi **16**(63), 1078 (2021)