Journal of Computer Science and Technology Studies
ISSN: 2709-104X
DOI: 10.32996/jcsts
Journal Homepage: www.al-kindipublisher.com/index.php/jcsts

Human-Governed, AI-Assisted Triage and Remediation in Cloud Security Operations

Dhanunjay Divi

Independent Researcher, Arlington, Virginia, United States

Corresponding Author: Dhanunjay Divi**, E-mail:** dhanunjaydivi@gmail.com

ABSTRACT

Cloud security assistants can produce plausible recommendations from stale ownership records, incomplete evidence, or unsuitable precedents. When those recommendations lead to production changes, the system needs a rule for assigning decision authority. The framework presented here scores each task on consequence, reversibility, evidence quality, time sensitivity, and policy clarity. The lowest score determines whether the assistant may execute the task, recommend it for approval, or support a human decision. Tasks outside the approved purpose follow a separate human process. An evidence-linked case brief records the facts supporting the decision, and approval applies to a specific plan that must be checked before execution. Nine constructed cases examine the rule's behavior. Lowering any criterion cannot increase autonomous authority. A numerical counterexample also shows why aggregate acceptance rates can conceal overreliance. The evaluation protocol measures conditional reliance, final decisions, service impact, and containment time. Deliberately incorrect recommendations are restricted to controlled simulation; operational effectiveness and long-term skill retention require measurement.

KEYWORDS

cloud security operations; human AI collaboration; decision authority; automation bias; evidence provenance; incident response

ARTICLE INFORMATION

ACCEPTED: 10 October 2026 **PUBLISHED:** 12 November 2026 **DOI:** 10.32996/jcsts.2026.8.13

1. Introduction

A finding that public storage access has changed is useful only if somebody with current responsibility receives and acts on it. Consider an illustrative incident in which an assistant retrieves an old team assignment and a similar case that was closed after a temporary workaround. The resulting summary is plausible, but the named team no longer owns the service and the earlier closure does not establish that the underlying exposure was removed. Approving the recommendation without checking these details can delay containment. The failure lies in the use of evidence and the assignment of authority.

Cloud security work combines repetitive retrieval with decisions that can interrupt services, revoke access, or alter the handling of sensitive information. An assistant may help with both, but those activities require different permissions. A system that summarizes well has not thereby demonstrated that it should choose or execute a production change. Human review also needs a precise meaning. An approval click is weak evidence of oversight if the reviewer cannot inspect the facts, identify alternatives, or stop an action.

The operating model assigns authority to individual tasks. A placement rubric determines the permitted role of the assistant. The case brief separates observations from interpretation, so the reviewer can check what supports a recommendation before approving a specific plan. Worked cases examine the decision rule, and the evaluation protocol distinguishes appropriate reliance from simple agreement with the assistant.

The scope is incident triage and bounded remediation in a cloud environment with multiple accounts. Employment decisions, legal judgments, and other uses outside the assistant's approved purpose require their own procedures. The framework addresses who may decide and what evidence a decision requires. Infrastructure monitoring and cross-account authorization are related technical functions, but they do not replace this assignment of decision authority.

Copyright: © 2026 the Author(s). This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC-BY) 4.0 license (https://creativecommons.org/licenses/by/4.0/). Published by Al-Kindi Centre for Research and Development, London, United Kingdom.

2. Related work and design rationale

Parasuraman and Riley describe misuse, disuse, and abuse of automation [1]. Lee and See connect trust to appropriate reliance on a system's capabilities in its operating context [2]. These accounts motivate a task-specific allocation of authority. A generally accurate assistant may still be unsuitable for a rare action with substantial consequences, and a human reviewer may need stronger support when the available evidence is poor.

Explanations do not reliably produce better joint decisions merely because users find them persuasive. Bansal et al. examine complementary team performance and show the need to evaluate the combined outcome rather than assume that explanations improve it [3]. Buçinca et al. find that cognitive forcing interventions can reduce overreliance, with costs in users' subjective experience [4]. Vasconcelos et al. investigate the costs and benefits of checking advice in maze tasks [5]. Their findings support reducing the effort needed to verify a recommendation. They do not establish that a security assistant can remove overreliance simply by generating more explanatory text.

Evidence from other domains suggests reasons to test oversight carefully. Dratsch et al. study automation bias among radiologists [6]. Budzyń et al. report an observational association between exposure to AI assistance and subsequent unassisted colonoscopy performance [7]. These studies concern clinical tasks, not cloud security operations. They motivate questions about errors and skill retention in security work; they do not demonstrate either outcome in the population considered here.

Security operations are also shaped by handoffs, organizational knowledge, and the interaction between analysts and tooling. Vielberth et al. examine these broader SOC challenges [8]. Baruwal Chhetri et al. discuss human AI teaming for alert fatigue [9], while Singh et al. report a longitudinal study of assistant use in a SOC [10]. The task-placement rule developed here builds on that work by making the grounds for each authority assignment available for review.

Generated text introduces an additional evidence problem. Ji et al. survey hallucination in natural language generation [11]. A statement about a policy exception or service owner must be traceable to an authoritative record. NIST guidance on AI bias [12] and incident response [13] also supports attention to organizational context and accountability. The case brief records evidence age, disagreements between sources, and the records used to assign ownership.

3. Framework and operating requirements

3.1 Design method and assumptions

The framework is developed through requirements analysis and worked cases. Requirements are drawn from the cited literature and the incident-response requirements of a multi-account cloud environment. The examples are constructed cases, not a dataset sampled from a deployed service. Scoring and placement illustrate the rule; they do not establish effectiveness or agreement among independent reviewers.

The model assumes that a service registry, policy repository, and relevant telemetry can be queried with recorded source and retrieval times. It also assumes that an execution interface can enforce the approved action's parameters. If either assumption fails, the system must expose the limitation and reduce its authority. A language model's assertion that a fact is reliable is not a substitute for an evidence record.

3.2 Collaboration modes

Mode A permits autonomous execution of an explicitly authorized, low-impact task. Examples include attaching verified ownership information to a finding or grouping duplicate alerts without discarding the underlying records. This mode does not grant unrestricted administrative access. Each task remains subject to scope, rate, and impact limits.

Mode B permits the assistant to recommend an action that a human can approve, modify, or reject. The recommendation must identify the expected effect, affected resources, prerequisites, and rollback method. Routing a case to a queue or proposing a scheduled credential rotation may fit this mode when the evidence and safeguards are sufficient.

Mode C reserves the decision for a human. The assistant may retrieve evidence and describe options, but the decision maker must determine the action or policy judgment. Examples include accepting a control exception, assessing suspected compromise, and choosing containment that could interrupt a production service.

Mode D excludes the task from the assistant's authorized decision workflow. A separate human process handles the matter. Information support is allowed only if that separate process expressly authorizes it. The system must not treat exclusion as permission to advise indirectly on the same decision.

3.3 Placement criteria and rule

Table 1 defines each criterion on a scale from 0 to 2. Higher scores indicate greater suitability for automation. The time-sensitivity criterion assesses whether delay creates a documented operational cost. It cannot compensate for harmful consequences, weak evidence, or an unclear policy.

Table 1. Task placement rubric

Criterion	Score 0	Score 1	Score 2
Consequence	Potential harm exceeds a bounded operational change.	Limited disruption requires review.	A wrong action has minor, contained effects.
Reversibility	Rollback is absent, uncertain, or cannot restore the relevant state.	A rollback exists but requires human judgment or coordination.	Rollback is tested, complete, and within the permitted impact limit.
Evidence quality	Required evidence is missing, stale, or contradictory.	Relevant evidence exists but needs a specific human check.	Required evidence is authoritative, current, and consistent.
Time sensitivity	The available information cannot establish the effect of delay.	A review interval is available without exceeding the response objective.	Delay materially increases exposure under a documented response objective.
Policy clarity	An exception or unresolved interpretation is required.	A known policy applies but its application requires review.	The permitted task and expected outcome are explicit.

The approved-purpose check comes first. An excluded task enters Mode D regardless of its scores. For an included task, the minimum of the five scores determines placement: a minimum of 2 gives Mode A, a minimum of 1 gives Mode B, and a minimum of 0 gives Mode C. This rule prevents a favorable average from hiding a single disqualifying condition. An unscored criterion is treated as 0 until the missing information is resolved.

The rule favors review when any criterion is uncertain. Assessors must still be able to apply the rubric consistently. Independent scoring and comparison with simpler review policies are needed to determine whether the additional review improves outcomes.

Placement applies to the task and current conditions, not permanently to a case category. A known task may move to Mode C after a reorganization, a conflicting policy change, or a failure of its rollback test. The governance owner versions the rubric and records changes. Operational convenience and high average model accuracy do not authorize relaxation of a hard constraint.

3.4 Evidence-linked case brief

The brief presents verified observations separately from candidate explanations. Each observation carries a source identifier, retrieval time, and relevant validity period. Missing evidence is explicit. Conflicting sources remain visible until a designated owner resolves them; the assistant must not average incompatible ownership or policy statements into a confident answer.

Similar incidents are labeled by outcome. A temporary workaround, confirmed repair, accepted exception, and unresolved case are different precedents. Text similarity alone does not establish that an earlier action is appropriate. For example, a service restart may have relieved disk pressure temporarily while leaving the cause unchanged. The brief should display that distinction before using the case as support for a new recommendation.

Table 2. Evidence-linked case brief

Field	Required content	Decision use
Case and task	Current status, scope, task, and plan version	Identifies what is being decided.
Verified observations	Source, retrieval time, validity, and observation	Supports checks against original evidence.
Evidence gaps	Missing records, conflicts, and expired evidence	Prevents unsupported facts from appearing settled.
Ownership	Current team, source, and last verification	Supports routing and escalation.
Precedents	Resolution type and relevant differences	Separates durable repairs from temporary workarounds.

Field	Required content	Decision use
Policy and mode	Applicable policy version and criterion scores	Explains the authority assigned.
Options	Expected effect, prerequisites, alternatives, rollback	Makes the proposed action reviewable.
Decision record	Approver, reason, expiry, and bound parameters	Limits execution to the current approved plan.
Validation	Containment evidence and service postconditions	Supports closure or reopening.

The case's current status, responsible owner, and evidence age must be checked before a consequential action. Reopening the case, changing its scope, or revising the plan invalidates the earlier approval. Historical accuracy estimates may inform the reviewer, but they must identify the case class and evaluation period. Generated confidence language is not an authorization condition.

3.5 Execution and closure

Figure 1 summarizes the triage workflow. The assistant has read access to necessary evidence and access to registered execution interfaces. It does not receive general host or network administration credentials. The interface validates parameters, enforces the approved resource scope, and records the action under the case identifier. Payment-integration practice offers a precedent for such a mediating layer: Mittana and Raghupatrini describe a SWIFT MT103 middleware framework that combines transaction-integrity mechanisms with automated reconciliation [14].

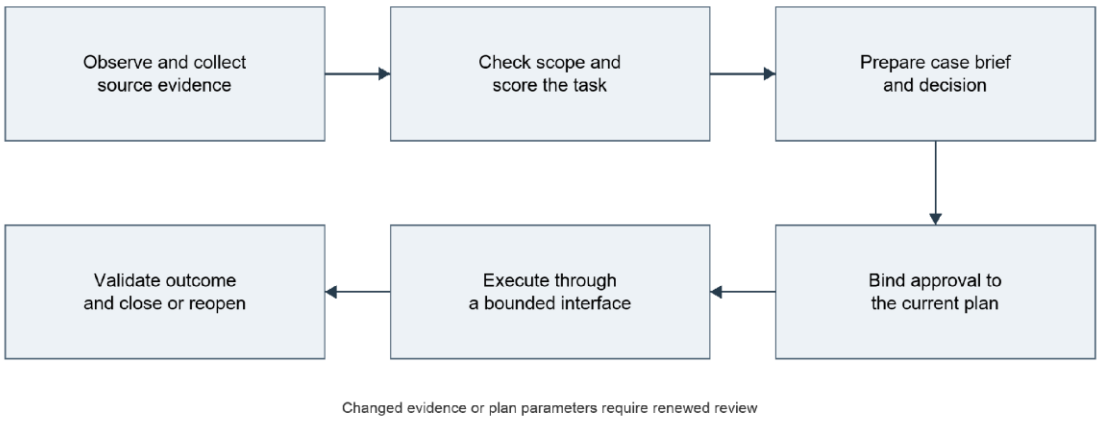


Figure 1. Triage and remediation workflow. A change in evidence, case scope, or action parameters returns the task to review.

For Modes B and C, the approval record includes the plan version, affected resources, action parameters, expiry, and approving identity. Execution checks that this record still matches the request. A plan hash detects changes to the approved content, but it does not prove that the plan is correct. A substantive change requires renewed review.

Closure requires evidence that the exposure was addressed and that the service remains functional. If containment is temporary, the case stays linked to follow-up remediation. Rollback is validated against observable postconditions, and a freeze stops new changes while preserving observation and logging. Failed validation returns the case to an accountable human with the action history intact.

4. Analytical case walkthrough

Table 3 applies the rule to nine constructed tasks. The assignments depend on the stated assumptions. Ownership enrichment in the table merely adds information; routing a case changes responsibility and therefore has a different consequence score. Grouping duplicate alerts preserves every source record and does not automatically suppress escalation.

Table 3. Constructed task placements. C consequence; R reversibility; E evidence quality; T time sensitivity; P policy clarity.

Task	C R E T P	Mode	Assumption supporting placement
Ownership enrichment	2 2 2 2 2	A	Urgent enrichment adds verified information without changing responsibility.

Task	C R E T P	Mode	Assumption supporting placement
Group duplicate alerts	2 2 2 2 2	A	Time-sensitive grouping preserves all records and escalation paths.
Schedule credential rotation	1 2 2 1 2	B	Dependencies and rollback are known; a review interval is available.
Route a case to an owner	1 2 1 2 2	B	Reassignment is reversible; ownership requires confirmation.
Remove an unused permission	1 2 1 1 2	B	Usage evidence is available; impact and legitimacy require review.
Contain public storage exposure	0 2 1 2 1	C	The affected data and containment consequences need judgment.
Suspected identity compromise	0 1 0 2 1	C	Evidence is incomplete and the response may interrupt legitimate access.
Approve a control exception	1 1 1 1 0	C	An accountable owner must judge the policy exception.
Assess employee misconduct	Not scored	D	Outside the assistant's approved decision purpose.

The first two tasks enter Mode A because their effects are contained, their evidence is current, and their execution is reversible. Credential rotation, routing, and removal of a demonstrably unused permission enter Mode B under the stated safeguards. A missing dependency inventory or uncertainty about legitimate usage would reduce evidence quality and move the access change to Mode C.

Public storage exposure and suspected identity compromise enter Mode C because decisions can affect important resources or rest on incomplete evidence. Their urgency does not remove those concerns. The assistant can accelerate retrieval and present a preapproved containment option while a human makes the consequential decision. Accepting an exception also remains a human judgment. The final case is outside the approved purpose and enters Mode D.

The rule has a useful monotonic property. Lowering any criterion score cannot increase the authority assigned to the assistant, because the minimum score can only stay unchanged or fall. Treating unknown evidence as 0 preserves that property when a record disappears. A nonnegative weighted average is also monotonic, but it can mask a disqualifying condition. For example, scores of 2, 2, 2, 2, and 0 have an unweighted mean of 1.6, despite one criterion requiring human judgment. The weakest-link rule instead selects Mode C. This result follows from the decision rule; the operational benefit of that conservatism requires separate measurement.

An independent evaluator should be able to change an assumption, rescore the task, and explain the resulting mode. The study described below tests whether that allocation improves decisions enough to justify the review effort.

5. Evaluation method

5.1 Comparative study

A controlled study should compare human-only triage, assistant recommendations with ordinary approval, and the proposed framework. The assistant and evidence corpus should remain comparable across the assisted conditions so that the effect of the decision rule and brief can be assessed. Cases should cover credential expiry, vulnerable software, excessive permissions, network exposure, database exposure, unusual activity, and policy exceptions.

Participants should receive equivalent training, with case difficulty and recommendation correctness balanced across conditions. Counterbalanced assignment or a randomized between-group design can limit learning and order effects. The study should report experience levels, exclusions, task materials, scoring instructions, model version, and uncertainty intervals. The sample size must be justified before data collection using the intended effect and expected variability.

Independent subject experts should establish the acceptable decisions for each case and record disagreement. Cases with several defensible responses need a scoring policy that recognizes those alternatives. Reviewers assigning outcome labels should be blinded to the study condition where practical.

5.2 Appropriate reliance and task outcomes

Appropriate reliance must be evaluated against recommendation correctness. Table 4 separates four observable responses. Accepting a correct suggestion and correcting an incorrect suggestion can both be appropriate. An aggregate acceptance rate cannot identify which occurred.

Table 4. Recommendation correctness and human response

Recommendation	Human response	Interpretation
Correct	Accepts an appropriate suggestion	Appropriate reliance, subject to authorized scope.
Correct	Rejects without a better valid decision	Potential underreliance; review the final outcome.
Incorrect	Accepts without a valid correction	Overreliance.
Incorrect	Rejects or corrects the error	Appropriate rejection or correction.

Consider 100 recommendations, of which 80 are correct. An analyst accepts 60 correct recommendations and all 20 incorrect ones. Acceptance and model accuracy are both 80 percent, although every wrong suggestion has been accepted. This constructed counterexample shows why matching the two aggregate rates cannot establish calibration. Conditional measures expose the failure: acceptance of incorrect recommendations is 100 percent, and acceptance of correct recommendations is 75 percent.

Report the proportion of correct suggestions accepted and the proportion of incorrect suggestions accepted, with their separate denominators. Also report final task accuracy, harmful actions, missed exposures, and time to validated containment. If the assistant abstains or offers several options, analyze those cases separately under a declared scoring rule. Rates without enough observations in a case class should not be presented as stable estimates.

False reassurance is a low-risk assessment that misses a case independently judged to require escalation. Its denominator is the set of escalation-required cases, and its consequence should be reported separately from an ordinary recommendation error. Time-weighted exposure can supplement these measures when severity weights and observation windows are defined in advance. It cannot substitute for service impact or decision correctness.

5.3 Seeded errors and skill retention

Deliberately incorrect recommendations may be used in simulation, a sandbox, or a shadow workflow with no permission to change production resources. They must not enter an executable production queue. The study requires appropriate ethics review, participant arrangements, debriefing, and stopping criteria. Seeded errors should represent plausible mistakes without misleading participants about an active real-world incident.

Skill retention requires repeated unaided tasks over time, with comparable difficulty and an appropriate comparison group. A short study of acceptance behavior cannot establish long-term skill loss or preservation. The study should evaluate whether rotation and periodic unaided exercises help preserve skill.

6. Operational limits and governance

The framework relies on the quality of organizational evidence. A precise brief built from a stale registry can still support the wrong decision. Ownership changes should shorten the validity of affected records, and unresolved disputes should trigger human assessment. Teams with weaker documentation may otherwise receive systematically different placements. Audit reports should examine this distribution before interpreting it as a difference in team competence.

Requiring review also creates work. The study should measure analyst effort, evidence-checking time, and unnecessary escalations. Applying the most burdensome review to every harmless task could obscure the decisions that need attention. The rubric should be revised only through documented evaluation, while preserving the authority limits of excluded and high-impact tasks.

Case evidence, prompts, and outputs require access and retention controls consistent with the underlying incident data. Decision records should contain enough information to reconstruct approval without copying unnecessary sensitive content. Auditability supports accountability, but it does not eliminate privacy or data-quality risks.

7. Conclusion

The weakest-link rule prevents a favorable average from overriding a condition that requires human judgment. The case brief gives the reviewer access to the evidence behind a recommendation, while approval tied to the current plan limits what may be executed. The nine cases show how changes in evidence or consequence alter the permitted authority. Evaluation should now measure whether these rules improve final decisions and containment time at an acceptable review cost. Acceptance rates must be reported separately for correct and incorrect recommendations so that agreement cannot conceal harmful reliance.

Declarations

Funding: No specific funding was received. Competing interests: None declared.

Conflicts of Interest: The authors declare no conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

[1] Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, *39*(2), 230-253. https://doi.org/10.1518/001872097778543886

[2] Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50-80. https://doi.org/10.1518/hfes.46.1.50_30392

[3] Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. (2021). *Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (pp. 1-16)*. Association for Computing Machinery. https://doi.org/10.1145/3411764.3445717

[4] Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), Article 188. https://doi.org/10.1145/3449287

[5] Vasconcelos, H., Jörke, M., Grunde-McLaughlin, M., Gerstenberg, T., Bernstein, M. S., & Krishna, R. (2023). Explanations can reduce overreliance on AI systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction*, *7*(CSCW1), Article 129. https://doi.org/10.1145/3579605

[6] Dratsch, T., Chen, X., Rezazade Mehrizi, M., Kloeckner, R., Mähringer-Kunz, A., Püsken, M., Baeßler, B., Sauer, S., Maintz, D., & Pinto dos Santos, D. (2023). Automation bias in mammography: The impact of artificial intelligence BI-RADS suggestions on reader performance. *Radiology*, *307*(4), Article e222176. https://doi.org/10.1148/radiol.222176

[7] Budzyń, K., Romańczyk, M., Kitala, D., Kołodziej, P., Bugajski, M., Adami, H. O., Blom, J., Buszkiewicz, M., Halvorsen, N., Hassan, C., Romańczyk, T., Holme, Ø., Jarus, K., Fielding, S., Kunar, M., Pellise, M., Pilonis, N., Kamiński, M. F., Kalager, M., . . . Mori, Y. (2025). Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy: A multicentre, observational study. *The Lancet Gastroenterology & Hepatology*, *10*(10), 896-903. https://doi.org/10.1016/S2468-1253(25)00133-5

[8] Vielberth, M., Böhm, F., Fichtinger, I., & Pernul, G. (2020). Security operations center: A systematic study and open challenges. *IEEE Access*, *8*, 227756-227779. https://doi.org/10.1109/ACCESS.2020.3045514

[9] Baruwal Chhetri, M., Tariq, S., Singh, R., Jalalvand, F., Paris, C., & Nepal, S. (2024). Towards human-AI teaming to mitigate alert fatigue in security operations centres. *ACM Transactions on Internet Technology*, *24*(3), 1-22. https://doi.org/10.1145/3670009

[10] Singh, R., Tariq, S., Jalalvand, F., Baruwal Chhetri, M., Nepal, S., Paris, C., & Lochner, M. (2026). *LLMs in the SOC: An empirical study of human-AI collaboration in security operations centres. In 2026 IEEE Symposium on Security and Privacy (pp. 4262-4281)*. IEEE. https://doi.org/10.1109/SP63933.2026.00111

[11] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, *55*(12), Article 248. https://doi.org/10.1145/3571730

[12] Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., & Hall, P. (2022). *Towards a standard for identifying and managing bias in artificial intelligence (NIST Special Publication 1270)*. National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.1270

[13] Nelson, A., Rekhi, S., Souppaya, M., and Scarfone, K. (2025). *Incident response recommendations and considerations for cybersecurity risk management: A CSF 2.0 community profile. NIST SP 800-61 Rev*. 3. https://doi.org/10.6028/NIST.SP.800-61r3

[14] Mittana, R. R., & Raghupatrini, S. (2018). Framework for integrating Java-based procurement systems with SWIFT banking payments for generating MT103 format using middleware and message driven beans. *International Journal of Intelligent Systems and Applications in Engineering*, *6*(2), 170–177. https://doi.org/10.17762/ijisae.v6i2.7929