Students are to answer 1 of the 2 cases presented. See below for essay length and formatting requirements. The included article gives relevant background information. Students can use any resource given in class or found online. Online resources should be evidence-based on peer-reviewed sources.

Case 1: Shared Mental Models in a Human–AI Security Operations Centre

A national cyber security operations centre (SOC) has deployed an AI-based decision support system that correlates logs, prioritises alerts, and recommends response actions in major incidents. After several serious attacks, reviews show that analysts, incident managers, and the AI system often operate with incompatible understandings of what is happening, what the priorities are, and who should do what. Some analysts treat the AI's risk scores as final decisions, others largely ignore them; incident commanders complain that they cannot see "what the AI thinks is going on", while the AI frequently recommends actions that do not match existing playbooks or the commander's intentions. Debriefs repeatedly reveal automation surprises, conflicting interpretations of the same data, and confusion about goals, roles, and plans. Management believes that the core problem is not only trust in automation, but misaligned shared mental models and shared situation awareness within the human–AI team.

Drawing on the attached journal article on situation awareness and shared mental models in human–AI systems, analyse the SOC's breakdowns, explain how gaps in individual and team situation awareness contribute to these problems, and show how more aligned shared mental models could improve performance. Propose and justify concrete changes to interface design, team roles and procedures, and training or exercises that could help the human analysts and the AI system develop, maintain, and update shared mental models during dynamic incidents. Critically evaluate the likely benefits and limitations of your proposal, including possible unintended consequences and outline how you would empirically assess whether shared situation awareness and team performance have actually improved.

Endsley, M. R. (2023). Supporting Human-AI Teams: Transparency, explainability, and situation awareness. *Computers in Human Behavior*, *140*, 107574.

Case 2:

Exam Task: Organisational Culture, Mental Health, and Insider Threats in a Technology Firm

A technology company providing cloud services has experienced several near-miss security incidents involving privileged employees. In one case, a system administrator copied large amounts of production data to a personal device shortly before resigning. An internal investigation found no malicious intent but revealed high levels of stress, burnout, and perceived injustice among technical staff. Employee surveys show low psychological safety, weak leadership support for mental health, and a belief that "security is mostly about tools, not people." At the same time, the security team has started to view dissatisfied employees primarily as potential insider threats, which has damaged trust even further.

The board has concluded that the organisation's culture and leadership practices may be increasing both psychosocial risk and security risk. You are brought in as an external consultant to design a comprehensive, human-centred insider threat mitigation strategy that does not rely solely on surveillance or technical controls.

Your task is to propose an integrated programme that addresses organisational culture, leadership behaviour, mental health, and security practices as interconnected elements. Use relevant theories and empirical research on insider threats, occupational stress and burnout, organisational justice, and safety/security climate. Outline concrete interventions at the individual, team, and organisational levels and critically evaluate ethical tensions between privacy and monitoring, and explain how you would assess the effectiveness and potential unintended effects of your proposed programme over time.

Khan, N., J. Houghton, R., & Sharples, S. (2022). Understanding factors that influence unintentional insider threat: a framework to counteract unintentional risks. *Cognition, Technology & Work*, *24*(3), 393-421.

Formatting of Essay

The essay can be **3000 ± 10% words** in main body (abstract and references do not count) in 12-point Times New Roman font. Your essay should be typed and double-spaced on standard-sized paper (A4) with 1" margins (standard in Word Document) on all sides. Include a page header (also known as the "running head") at the top of every page. The running head is a shortened version of your paper's title and cannot exceed 50 characters including spacing and punctuation. Not meeting this requirement can result in a lower grade. Word count does not include pictures with no text. Tables will count towards the final word count. If pictures are used for tables, they will count as 500 words. The essay shall have a title page with Title and word count listed, abstract, main body, and references. This essay is an academic text.

See https://taltech.ee/en/formatting-guidelines for more information

For this essay, APA 7th is to be used (see: https://apastyle.apa.org/style-grammar-guidelines/references/examples)

Other resources for APA 7th (I personally use this one): https://owl.purdue.edu/owl/research_and_citation/apa_style/apa_style_introduction.html

Grading

See https://taltech.ee/en/grading-system for Taltech Grading policy

§ 14. Assessment of academic performance

(1) The methods and criteria of assessment defined in syllabus shall be available to students before the commencement of studies and they must not be changed during a semester. The assessment methods define the manner of attesting the acquisition of knowledge and skills (e.g. an oral or written examination, pass/fail assessment, an essay, a report, group work, a questionnaire). If various methods are used for the assessment of learning outcomes, their relevant weights in determining the final grade shall be specified in the syllabus. An assessment criterion shall specify the expected level and scope of knowledge which can be proved by the assessment methods.

(3) **In case of graded assessment, the achievement of learning outcomes is assessed based on the following scale**:

A (5) – "excellent" – outstanding and particularly profound achievement of learning outcomes, along with creativity and consummate proficiency in applying skills and knowledge;

B (4) – "very good" – very good achievement of learning outcomes, along with proficiency in applying skills and knowledge in a targeted and creative manner. Some details of knowledge and skills may exhibit errors which are neither substantive nor serious;

C (3) – "good" – good achievement of learning outcomes, along with proficiency in applying skills and knowledge in a relevant manner. A certain imprecision and uncertainty are apparent in the depth and detail of knowledge and skills;

D (2) – "satisfactory" – sufficient achievement of learning outcomes, along with application of knowledge and skills in a typical manner; in atypical situations both, uncertainty as well as lack of knowledge and skills are apparent.

E (1) – "poor" – minimum acceptable achievement of the most important learning outcomes along with limited application of knowledge and skills in typical situations; in atypical situations both, considerable uncertainty as well as lack of knowledge and skills are apparent; F (0) – "failed" – achievement in knowledge and skills below the minimum standard.

Οι φοιτητές καλούνται να επιλέξουν και να απαντήσουν σε μία από τις δύο διαθέσιμες περιπτώσεις. Παρακάτω θα βρείτε τις απαιτήσεις για το μέγεθος και τη μορφοποίηση του δοκιμίου. Το άρθρο που παρέχεται περιέχει σημαντικές πληροφορίες για το υπόβαθρο του θέματος. Μπορείτε να χρησιμοποιήσετε οποιοδήποτε υλικό έχει δοθεί στο μάθημα ή να αναζητήσετε αξιόπιστες πηγές στο διαδίκτυο, με την προϋπόθεση ότι βασίζονται σε επιστημονικά τεκμηριωμένες και αξιολογημένες δημοσιεύσεις.

Περίπτωση 1: Κοινά Νοητικά Μοντέλα σε ένα Κέντρο Επιχειρήσεων Ασφαλείας Ανθρώπου–AI

Ένα εθνικό κέντρο επιχειρήσεων κυβερνοασφάλειας (SOC) έχει υιοθετήσει ένα σύστημα υποστήριξης αποφάσεων βασισμένο στην τεχνητή νοημοσύνη, το οποίο συσχετίζει αρχεία, δίνει προτεραιότητα στις ειδοποιήσεις και προτείνει ενέργειες αντιμετώπισης σε σοβαρά περιστατικά. Μετά από αρκετές σοβαρές επιθέσεις, οι αξιολογήσεις δείχνουν ότι οι αναλυτές, οι διαχειριστές περιστατικών και το σύστημα τεχνητής νοημοσύνης συχνά λειτουργούν με διαφορετικές αντιλήψεις σχετικά με το τι συμβαίνει, ποιες είναι οι προτεραιότητες και ποιος πρέπει να κάνει τι. Κάποιοι αναλυτές θεωρούν τις βαθμολογίες κινδύνου της ΤΝ ως τελικές αποφάσεις, άλλοι τις αγνοούν σχεδόν εντελώς· οι επικεφαλής περιστατικών παραπονιούνται ότι δεν μπορούν να καταλάβουν «τι πιστεύει η ΤΝ ότι συμβαίνει», ενώ το σύστημα συχνά προτείνει ενέργειες που δεν ταιριάζουν με τα υπάρχοντα εγχειρίδια ή τις προθέσεις του επικεφαλής. Οι απολογισμοί φανερώνουν συνεχώς εκπλήξεις από την αυτοματοποίηση, αντιφατικές ερμηνείες των ίδιων δεδομένων και σύγχυση σχετικά με στόχους, ρόλους και σχέδια. Η διοίκηση πιστεύει ότι το βασικό πρόβλημα δεν είναι μόνο η εμπιστοσύνη στην αυτοματοποίηση, αλλά και η ελλιπής ευθυγράμμιση των κοινών νοητικών μοντέλων και της συλλογικής επίγνωσης κατάστασης στην ομάδα ανθρώπων–ΤΝ.

Αξιοποιώντας το συνημμένο επιστημονικό άρθρο για την επίγνωση κατάστασης και τα κοινά νοητικά μοντέλα σε συστήματα ανθρώπου–τεχνητής νοημοσύνης, αναλύστε τις δυσλειτουργίες του SOC, εξηγήστε πώς τα κενά στην ατομική και ομαδική επίγνωση κατάστασης συμβάλλουν σε αυτά τα προβλήματα και δείξτε πώς πιο ευθυγραμμισμένα κοινά νοητικά μοντέλα θα μπορούσαν να βελτιώσουν την απόδοση. Προτείνετε και τεκμηριώστε συγκεκριμένες αλλαγές στο σχεδιασμό της διεπαφής, στους ρόλους και τις διαδικασίες της ομάδας, καθώς και στην εκπαίδευση ή τις ασκήσεις, που θα βοηθούσαν τους ανθρώπινους αναλυτές και το σύστημα ΤΝ να αναπτύξουν, να διατηρήσουν και να επικαιροποιούν κοινά νοητικά μοντέλα κατά τη διάρκεια δυναμικών περιστατικών. Αξιολογήστε κριτικά τα πιθανά οφέλη και τους περιορισμούς της πρότασής σας, συμπεριλαμβανομένων πιθανών ανεπιθύμητων συνεπειών, και περιγράψτε πώς θα αξιολογούσατε εμπειρικά εάν η συλλογική επίγνωση κατάστασης και η απόδοση της ομάδας έχουν πραγματικά βελτιωθεί.

Endsley, M. R. (2023). Υποστήριξη Ομάδων Ανθρώπου-ΤΝ: Διαφάνεια, εξηγησιμότητα και επίγνωση κατάστασης. *Computers in Human Behavior*, *140*, 107574.

Περίπτωση 2:

Εξεταστική Εργασία: Οργανωσιακή Κουλτούρα, Ψυχική Υγεία και Εσωτερικές Απειλές σε Εταιρεία Τεχνολογίας

Μια εταιρεία τεχνολογίας που παρέχει υπηρεσίες cloud ήρθε αντιμέτωπη με αρκετά περιστατικά ασφαλείας που παραλίγο να εξελιχθούν σε σοβαρά, με εμπλοκή προνομιούχων υπαλλήλων. Σε μία περίπτωση, διαχειριστής συστημάτων αντέγραψε μεγάλες ποσότητες παραγωγικών δεδομένων σε προσωπική του συσκευή λίγο πριν παραιτηθεί. Η εσωτερική έρευνα δεν εντόπισε κακόβουλη πρόθεση, αλλά ανέδειξε υψηλά επίπεδα άγχους, επαγγελματικής εξουθένωσης και αίσθησης αδικίας ανάμεσα στο τεχνικό προσωπικό. Οι απαντήσεις των εργαζομένων σε σχετικές έρευνες φανερώνουν χαμηλό αίσθημα ψυχολογικής ασφάλειας, ελλιπή στήριξη από τη διοίκηση σε θέματα ψυχικής υγείας και την αντίληψη ότι «η ασφάλεια είναι κυρίως θέμα εργαλείων, όχι ανθρώπων». Παράλληλα, η ομάδα ασφαλείας έχει αρχίσει να αντιμετωπίζει τους δυσαρεστημένους υπαλλήλους κυρίως ως πιθανούς εσωτερικούς κινδύνους, γεγονός που έχει διαβρώσει ακόμη περισσότερο την εμπιστοσύνη.

Το διοικητικό συμβούλιο κατέληξε στο συμπέρασμα ότι η εταιρική κουλτούρα και οι πρακτικές ηγεσίας ίσως ενισχύουν τόσο τους ψυχοκοινωνικούς όσο και τους κινδύνους ασφαλείας. Καλείστε, ως εξωτερικός σύμβουλος, να σχεδιάσετε μια ολοκληρωμένη και ανθρωποκεντρική στρατηγική αντιμετώπισης εσωτερικών απειλών, αποφεύγοντας να βασιστείτε αποκλειστικά σε παρακολούθηση ή τεχνικά μέτρα.

Η αποστολή σας είναι να προτείνετε ένα ολοκληρωμένο πρόγραμμα που ενσωματώνει τον οργανωσιακό πολιτισμό, τη συμπεριφορά ηγεσίας, την ψυχική υγεία και τις πρακτικές ασφάλειας ως αλληλένδετα στοιχεία. Αξιοποιήστε σύγχρονες θεωρίες και ερευνητικά ευρήματα σχετικά με εσωτερικές απειλές, εργασιακό άγχος και εξουθένωση, οργανωσιακή δικαιοσύνη και το κλίμα ασφαλείας/προστασίας. Προτείνετε συγκεκριμένες παρεμβάσεις σε ατομικό, ομαδικό και οργανωσιακό επίπεδο, αξιολογήστε κριτικά τα ηθικά διλήμματα μεταξύ ιδιωτικότητας και παρακολούθησης, και περιγράψτε πώς θα μετρήσετε την αποτελεσματικότητα και τις πιθανές ανεπιθύμητες συνέπειες του προγράμματος με την πάροδο του χρόνου.

Khan, N., J. Houghton, R., & Sharples, S. (2022). Κατανόηση των παραγόντων που επηρεάζουν την ακούσια εσωτερική απειλή: ένα πλαίσιο για την αντιμετώπιση ακούσιων κινδύνων. *Cognition, Technology & Work*, *24*(3), 393-421.

Μορφοποίηση της Εργασίας

Η εργασία μπορεί να έχει **3000 ± 10% λέξεις** στο κυρίως σώμα του κειμένου (η περίληψη και οι βιβλιογραφικές αναφορές δεν υπολογίζονται). Εφαρμόστε γραμματοσειρά Times New Roman, μέγεθος 12. Η εργασία πρέπει να είναι δακτυλογραφημένη, με διάστιχο 2, σε χαρτί A4 με περιθώρια 1" (τυπικές ρυθμίσεις Word) σε όλες τις πλευρές. Προσθέστε κεφαλίδα σε κάθε σελίδα (γνωστή και ως "τρέχουσα κεφαλίδα"). Η τρέχουσα κεφαλίδα είναι μια συντομευμένη εκδοχή του τίτλου της εργασίας σας και δεν πρέπει να ξεπερνά τους 50 χαρακτήρες, μαζί με τα κενά και τα σημεία στίξης. Η μη τήρηση αυτών των οδηγιών μπορεί να οδηγήσει σε χαμηλότερη βαθμολογία. Οι εικόνες χωρίς κείμενο δεν συμπεριλαμβάνονται στην καταμέτρηση των λέξεων. Οι πίνακες υπολογίζονται στον τελικό αριθμό λέξεων. Αν χρησιμοποιήσετε εικόνες ως πίνακες, θα μετράνε ως 500 λέξεις. Η εργασία πρέπει να περιλαμβάνει σελίδα τίτλου με τον τίτλο και τον αριθμό λέξεων, περίληψη, κυρίως σώμα και βιβλιογραφία. Η εργασία αυτή αποτελεί ακαδημαϊκό κείμενο.

Δείτε https://taltech.ee/en/formatting-guidelines για περισσότερες πληροφορίες

Για το συγκεκριμένο δοκίμιο, εφαρμόζεται η έκδοση 7 του APA

https://apastyle.apa.org/style-grammar-guidelines/references/examples)

Άλλες χρήσιμες πηγές για το APA 7η έκδοση (εγώ προσωπικά χρησιμοποιώ αυτήν):

https://owl.purdue.edu/owl/research_and_citation/apa_style/apa_style_introduction.html

Βαθμολόγηση

Δείτε https://taltech.ee/en/grading-system για την πολιτική βαθμολόγησης του Taltech

§ 14. Αξιολόγηση της ακαδημαϊκής επίδοσης

(1) Οι μέθοδοι και τα κριτήρια αξιολόγησης που περιγράφονται στο πρόγραμμα σπουδών πρέπει να είναι διαθέσιμα στους φοιτητές πριν την έναρξη των μαθημάτων και δεν επιτρέπεται να τροποποιηθούν κατά τη διάρκεια του εξαμήνου. Οι μέθοδοι αξιολόγησης καθορίζουν τον τρόπο πιστοποίησης της απόκτησης γνώσεων και δεξιοτήτων (π.χ. προφορικές ή γραπτές εξετάσεις, αξιολόγηση με επιτυχία/αποτυχία, έκθεση, αναφορά, ομαδική εργασία, ερωτηματολόγιο). Αν χρησιμοποιούνται διαφορετικές μέθοδοι για την αξιολόγηση των μαθησιακών αποτελεσμάτων, το πρόγραμμα σπουδών πρέπει να καθορίζει το σχετικό βάρος κάθε μεθόδου στην τελική βαθμολογία. Τα κριτήρια αξιολόγησης προσδιορίζουν το αναμενόμενο επίπεδο και εύρος των γνώσεων που μπορούν να αποδειχθούν μέσω των μεθόδων αξιολόγησης.

(3) **Σε περίπτωση αξιολόγησης με βαθμούς, η επίτευξη των μαθησιακών αποτελεσμάτων εκτιμάται σύμφωνα με την ακόλουθη κλίμακα**:

A (5) – «Άριστα» – Εξαιρετική και ιδιαίτερα εις βάθος επίτευξη μαθησιακών αποτελεσμάτων, με δημιουργικότητα και άριστη ικανότητα εφαρμογής γνώσεων και δεξιοτήτων.

B (4) – «Πολύ καλά» – Πολύ καλή επίτευξη μαθησιακών αποτελεσμάτων, με ικανότητα εφαρμογής γνώσεων και δεξιοτήτων με στόχευση και δημιουργικότητα. Ορισμένες λεπτομέρειες ενδέχεται να παρουσιάζουν λάθη που δεν είναι σοβαρά ή ουσιώδη.

C (3) – «Καλά» – Καλή επίτευξη μαθησιακών αποτελεσμάτων, με σχετική ικανότητα εφαρμογής γνώσεων και δεξιοτήτων. Υπάρχει κάποια ασάφεια ή αβεβαιότητα στο βάθος και τη λεπτομέρεια των γνώσεων και δεξιοτήτων.

D (2) – «Ικανοποιητικά» – Επαρκής επίτευξη μαθησιακών αποτελεσμάτων, με τυπική εφαρμογή γνώσεων και δεξιοτήτων. Σε μη συνηθισμένες καταστάσεις παρατηρείται αβεβαιότητα ή έλλειψη γνώσεων και δεξιοτήτων.

E (1) – «Ανεπαρκώς» – Ελάχιστα αποδεκτή επίτευξη των βασικών μαθησιακών αποτελεσμάτων, με περιορισμένη εφαρμογή γνώσεων και δεξιοτήτων σε συνήθεις περιπτώσεις. Σε ασυνήθιστες περιπτώσεις, διακρίνεται σημαντική αβεβαιότητα ή έλλειψη γνώσεων και δεξιοτήτων. F (0) – «Αποτυχία» – Επίδοση κάτω από το ελάχιστο αποδεκτό επίπεδο γνώσεων και δεξιοτήτων.