Unit 3 Seminar Preperation

1. **StereoSet: Measuring stereotypical bias in pretrained language models** *Moin Nadeem, Anna Bethke, Siva Reddy*. ACL / arXiv (2020/2021). This paper presents a large benchmark dataset (StereoSet) and evaluates popular pretrained language models for stereotypical biases (gender, profession, race, religion). Full paper and dataset/code are available (ACL proceedings / arXiv / GitHub). DOI: https://doi.org/10.18653/v1/2021.acl-long.416

2. **Factors affecting cybersecurity awareness: A qualitative study** *GrowingScience / authors listed in the PDF (2024)*: Qualitative research using focus groups or interviews to explore how people perceive cybersecurity, which internal and external factors influence behaviour, and the effectiveness of awareness programmes. The PDF is openly available under: growingscience.com

Evaluation Criteria	**Paper 1: Nadeem et al. (2021) – "StereoSet: Measuring stereotypical bias in pretrained language models"** ***(Quantitative / Experimental)***	**Paper 2: GrowingScience (2024) – "Factors affecting cybersecurity awareness: A qualitative study" (Qualitative / Thematic)**
Purpose, Problem, Objective	To build a large-scale benchmark ("StereoSet") to measure how pretrained language models (e.g., GPT-2, BERT) display social stereotypes across gender, race, and religion. Addresses fairness and bias in AI systems.	To explore why cybersecurity awareness training often fails and which personal or organisational factors influence user behaviour. Focuses on human elements rather than technical ones.
Contribution to the field	Contributes a *standardised measurement tool* for bias, enabling comparisons across models, key for ethical AI evaluation. Extends previous small-scale studies.	Contributes practical insight into how employees perceive cybersecurity messages, informing more effective training and culture-building strategies. Addresses human factors often ignored in technical studies.
Research Methodology	Quantitative, experimental benchmarking using structured datasets and statistical evaluation (bias scores, association tests).	Qualitative, exploratory study using semi-structured interviews and focus groups with thematic analysis.
Is the methodology appropriate?	Yes, objective, replicable method suitable for comparing bias metrics across systems. However, lacks context about *why* biases appear.	Yes, Interviews are appropriate for understanding attitudes and perceptions. However, limited sample size affects generalisability.
Data Collection & Analysis	Uses >17,000 crowd-sourced sentences; bias is measured via "association tasks." Statistical analysis is appropriate for the aim. Risk: benchmark data may not fully represent real-world usage.	Data gathered from professionals through focus groups; analysed with coding and theme identification. Provides depth and context but relies on interpretation.

Support for Claims / Evidence	Strong: empirical results and quantitative metrics clearly support conclusions. Tables and bias scores show transparent analysis.	Strong: participant quotes and thematic summaries support arguments, but claims are more interpretive. Limited triangulation.
Strengths	- Objective, reproducible metrics. - Public dataset encourages transparency and replication. - Fills a key measurement gap.	- Provides real-world insights. - Explains behavioural and organisational factors. - Highlights social context often missed by technical research.
Weaknesses / Limitations	- Benchmarking may oversimplify social bias. - Focused only on English-language data. - Quantitative method misses qualitative context (user impact).	- Limited sample size, mostly one region or organisation. - No quantitative validation of findings. - Subjective analysis risks researcher bias.
How Well Data Collection Supports Purpose	Well-suited for model comparison, but lacks longitudinal or behavioural data on bias outcomes.	Strong alignment, interviews reveal nuanced factors affecting cybersecurity habits.
Are Claims Supported by Evidence?	Quantitative results back claims about comparative bias between models.	Quotes and recurring themes back the conclusions; however, generalisation is limited.
Enhancement Suggestions	- Combine with qualitative interviews to explain *why* models show certain biases. - Add cross-lingual datasets to improve diversity. - Use real-world case studies.	- Add a follow-up survey to quantify how widespread certain attitudes are. - Include design to test behaviour change after training.
Fit Between Method and Question	Experimental design fits performance measurement aims.	Qualitative design fits exploratory human behaviour aims.
Evidence Quality and Credibility	High, open data, reproducibility, and statistical rigour.	Moderate, valid insights but smaller sample; relies on transparency and reflexivity for credibility.
How They Could Complement Each Other	Quantitative benchmarking gives scale; qualitative study adds human understanding. Together, they represent *mixed-methods strength*.	Same, qualitative insights can inform AI ethics; AI measurement approaches could inspire structured evaluation in cybersecurity training.
Personal Reflection	Reflects current concerns about bias in generative AI models, relevant for responsible computing practice.	Closely relates to workplace cybersecurity challenges, highlights the social side of digital risk management.
Overall Critical Evaluation	Strong empirical contribution but could expand interpretative depth and global inclusivity.	Strong contextual insight but limited in scope and generalisation.
How to Enhance the Work	Include interdisciplinary methods combining technical and ethical perspectives.	Combine qualitative findings with quantitative behavioural data for stronger evidence.

The best research design depends on the research question: quantitative for measurement and comparability; qualitative for exploring human experience. Both have complementary strengths