Introduction

Reference Slide:
https://canvas.lms.unimelb.edu.au/courses/117019/files/6826449?module_item_id=2703430

The aims of this tutorial are two-fold:

1. To **reinforce your understanding of** the relationship between (**contractual**) **trust** and **digital ethics**.
2. To critically **analyse a particular application** and **discuss** the **strengths** and **limitations** of it with respect to trust and digital ethics.

Task 1

Automated essay grading is the use of software to automatically assign (part of) a grade for a written essay. According to a 2019 research article (https://www.ijcai.org/Proceedings/2019/0879.pdf (Links to an external site.)

), the following factors are used to assign grades:

Dimension	**Description**
Grammar	Use of grammar
Usage	Use of prepositions, word usage
Mechanics	Spelling, punctuation, capitalisation
Style	Word choice, sentence structure
Relevance	Relevance of the content to the prompt
Organisation	How well the essay is structured
Development	Development of ideas with examples
Cohesion	Appropriate use of transition phrases
Coherence	Appropriate transitions between ideas
Thesis clarity	Clarify of the thesis/argument
Persuasiveness	Convincingness of the major argument

Imagine that the COMP90087 subject coordinators decide to use an automated essay grading tool that uses state-of-the-art in natural language processing technology. The tool uses the factors listed above to assign a grade.

NOTE: We will NOT be using an automated essay grader in COMP90087! :)

Discuss in your group:

Output

1. How would you react if you were told about the COMP90087 coordinators' decision?

Nhan: Would request a demo first to understand the new system of grading. Testing? Test to Extrinsic trust, see how it performs first, before the actual action, students might feel stress and concern to place their trust - place their eggs inside a black box, without seeing how it works first handed.

Jack: Would want to compare to a human marker's methodology - ensure that the training set that the AI was well tested against a set of *local* writing. Also would want to know that there were adequate methods of dispute resolutions if there were disagreements to the given mark. The possibility of marks being returned faster may be a benefit to the scheme.

Hong: I think the NLP application is not transparent enough for me to trust it. If the teaching team can give an explanation of the behind-the-scenes algorithms. I will trust it more.

Chris: I would request clear rubrics & programming parameters for the grading algorithm. While human preferences/intuition can be known/anticipated, inanimate algorithms are an unfamiliar territory for me. Furthermore, I would like to have the ability to submit for essay revaluation to a human marker if I find the marking done as unsatisfactory.

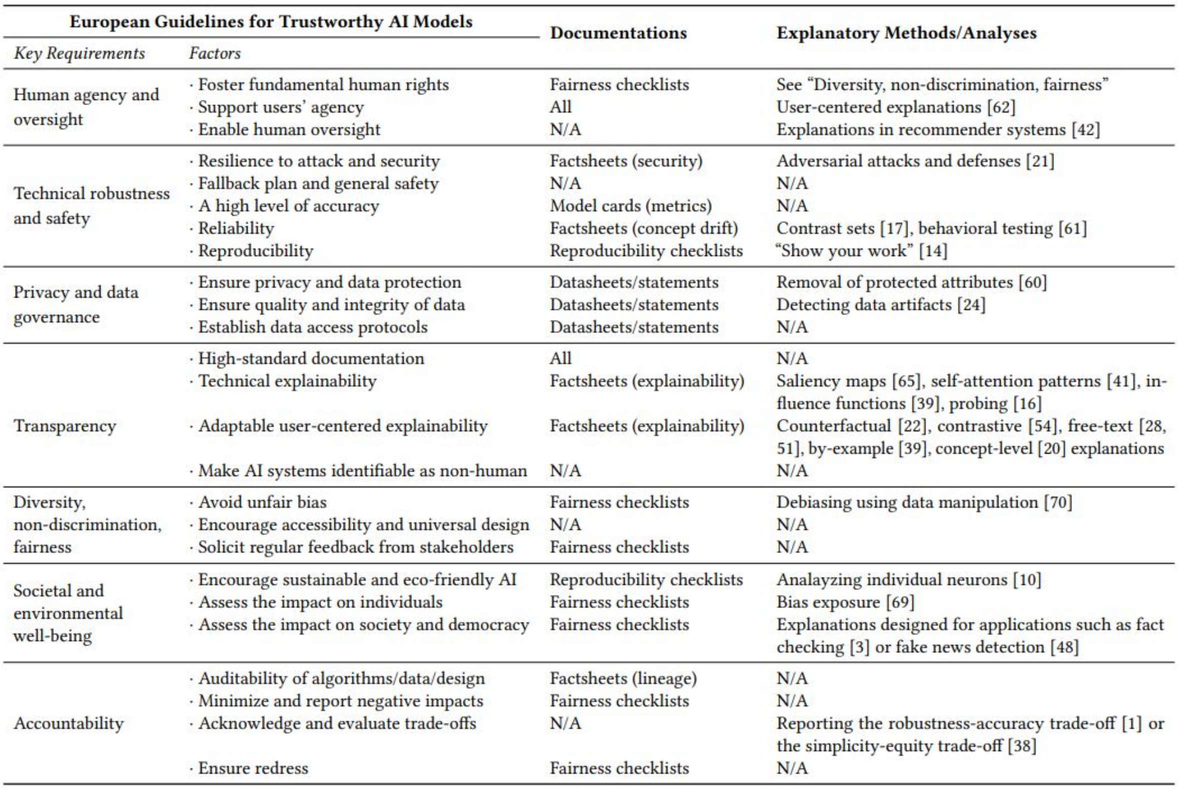
2. Using the model of contractual trust, **list three of the *factors*** from the table in the notes (reproduced below) that you think an automated essay grading tool would be unlikely to uphold?

1. Transparency
2. Accountability
3. Diversity, non-discrimination, fairness
4. Human agency and oversight

3. What are some of the **ethical risks** that could occur **if** these **factors/contracts are not upheld**? **Analyse these** against the concepts of use, misuse, abuse, and disuse of machines.

Individuals or groups might be able to manipulate that AI-grading tool, and give the person they favor or “buyers” the high mark? Furthermore, particular user groups might be disadvantaged if their write-up is non-standard as per training data-set. Similarly, the automated grading also poses risks for enabling fair grading of creative submissions as such instances will not be covered in the training set of the AI model.

In general, the teaching team has abused an NLP application that we (students) totally distrust. This unwarranted trust from the designer (the teaching team) makes use feel bad.



European Guidelines for Trustworthy AI Models		Documentations	Explanatory Methods/Analyses
Key Requirements	*Factors*		
Human agency and oversight	· Foster fundamental human rights	Fairness checklists	See "Diversity, non-discrimination, fairness"
	· Support users' agency	All	User-centered explanations [62]
	· Enable human oversight	N/A	Explanations in recommender systems [42]
Technical robustness and safety	· Resilience to attack and security	Factsheets (security)	Adversarial attacks and defenses [21]
	· Fallback plan and general safety	N/A	N/A
	· A high level of accuracy	Model cards (metrics)	N/A
	· Reliability	Factsheets (concept drift)	Contrast sets [17], behavioral testing [61]
	· Reproducibility	Reproducibility checklists	"Show your work" [14]
Privacy and data governance	· Ensure privacy and data protection	Datasheets/statements	Removal of protected attributes [60]
	· Ensure quality and integrity of data	Datasheets/statements	Detecting data artifacts [24]
	· Establish data access protocols	Datasheets/statements	N/A
Transparency	· High-standard documentation	All	N/A
	· Technical explainability	Factsheets (explainability)	Saliency maps [65], self-attention patterns [41], influence functions [39], probing [16]
	· Adaptable user-centered explainability	Factsheets (explainability)	Counterfactual [22], contrastive [54], free-text [28, 51], by-example [39], concept-level [20] explanations
	· Make AI systems identifiable as non-human	N/A	N/A
Diversity, non-discrimination, fairness	· Avoid unfair bias	Fairness checklists	Debiasing using data manipulation [70]
	· Encourage accessibility and universal design	N/A	N/A
	· Solicit regular feedback from stakeholders	Fairness checklists	N/A
Societal and environmental well-being	· Encourage sustainable and eco-friendly AI	Reproducibility checklists	Analayzing individual neurons [10]
	· Assess the impact on individuals	Fairness checklists	Bias exposure [69]
	· Assess the impact on society and democracy	Fairness checklists	Explanations designed for applications such as fact checking [3] or fake news detection [48]
Accountability	· Auditability of algorithms/data/design	Factsheets (lineage)	N/A
	· Minimize and report negative impacts	Fairness checklists	N/A
	· Acknowledge and evaluate trade-offs	N/A	Reporting the robustness-accuracy trade-off [1] or the simplicity-equity trade-off [38]
	· Ensure redress	Fairness checklists	N/A

Nominate a note taker to record your answers. Put your answers in the text of the forum discussion post, rather than as an attachment, to encourage others to read your notes.

Task 2

Your tutor will assign an additional task for Task 2. This is a 'holdout' task that we are not giving out in advance, as it will affect your answers to Task 1.

As a group, again document your answers for Task 2 once your tutor leads the discussion.

Why did/didn't you discuss the other stakeholders?

Being students, we approached the previous questions from a student's perspective; all issues raised previously were regarding the learning experience. Our lack of discussions regarding the other stakeholders may have stemmed from the group's large acknowledgement of the importance of grading throughout our higher education - and our ignorance to the actions “behind the scenes”.

Other stakeholders and their impact of using the automated grading system:

Subject coordinators: they distrust in the way of marking of their colleagues and they have a bias towards the NLP application that has not been tested carefully. So they are trying to deploy a system when it should not be used. The abuse of the NLP application causes a negative impact on us (students).

Students *outside of* the class: Future students who could be thinking of taking the subject might factor in the automated essay grading to decide whether to enrol. Especially if the essay constitutes a large portion of the grade, then it is very possible that students will start considering the risks (fairness, diversity, transparency, etc.).

University: There may be some ethical concerns, namely what will the money saved from the reduced reliance on teaching staff on marking. However, the university management staff will also need to consider potential ramifications with regulatory authorities such as the Department of Education, while implementing such an automated grading scheme.