AI Ethics and Sustainability

Ethical Considerations for Implementing AI Grading in High Ranking Higher Institutions (University of Southampton) in the UK

Module Code: MANG6585

Student ID: 37595717

Word count: 2200 words

Introduction

Grading has traditionally been understood as a human responsibility. The University of Southampton states that "academic judgement will always be necessary in order to arrive at a suitable grade" (University of Southampton, n.d.-b), while its feedback policy describes students as "partners in the process" (University of Southampton, 2023). This reflects the belief that assessment depends on trust, interpretation, and human understanding rather than mechanical calculation. Yet the assumption that human grading is naturally fair or consistent is difficult to defend. Markers experience fatigue, workload pressure, and implicit bias, all of which shape decision making. Southampton's moderation policy itself refers to situations involving staff sickness and reduced working capacity (University of Southampton, 2020), raising concerns about whether the current system achieves the consistency universities claim it does.

At the same time, AI grading introduces its own ethical problems. Sartor and Lagioia (2020) argue that a biased algorithm may still be fairer than an even more biased human decision maker, but this does not remove concerns about reliability and accountability. Large language models (LLMs) have been criticised for producing outputs without concern for truth (Hicks, Humphries and Slater, 2024), while Santos and Radanliev (2024) describe them as opaque systems where accountability becomes unclear. Without proper governance and stakeholder involvement (Miller, 2022), AI grading risks replacing one flawed system with another. This essay argues that AI grading at Southampton may only be ethically acceptable within a tightly governed hybrid model where meaningful human oversight remains central.

Why Human-Only Grading Appears Ethically Preferable

On the surface, keeping grading entirely in human hands seems ethically preferable. Assessment involves more than mechanically applying a rubric. It requires interpretation, contextual understanding, and judgement that cannot easily be reduced to fixed rules. Southampton's assessment framework recognises this directly, stating that students do not fit into "neat boxes" and that academic judgement is always necessary (University of Southampton, n.d.-a). A grade is therefore not simply a measurement but an interpretation of a student's understanding and argument.

Education also depends on relationships. Southampton's feedback policy frames students as "partners in the process" (University of Southampton, 2023), implying empathy, communication, and discretion. Students expect that if circumstances affect their work or something goes wrong, another human being will recognise the context and respond fairly. An algorithm cannot genuinely empathise or understand intent in the same way a person can.

Accountability is another major concern. Sparrow and Flenady (2025) argue that machines cannot function as moral agents and therefore cannot truly justify or take responsibility for their decisions. If an AI system awards an unfair grade, responsibility becomes difficult to locate. This issue is intensified by what Burrell (2016) describes as algorithmic opacity, where even developers struggle to explain how outputs are produced. Hicks, Humphries and Slater (2024) extend this criticism to LLMs specifically, arguing that they generate convincing language without any real concern for truth. Santos and Radanliev (2024) warn that these systems may create false authority by producing confident but inaccurate feedback.

The fear is therefore not simply that AI makes mistakes, but that universities may gradually replace meaningful educational judgement with automation because it is cheaper and more efficient. If assessment becomes primarily machine driven, education risks losing the human understanding that gives academic evaluation legitimacy.

However, this defence of human grading depends on the assumption that human markers themselves are objective and reliable, which becomes much harder to maintain under closer examination.

Questioning the Assumption of Human Objectivity

The argument for human-only grading often assumes that people are naturally fair and consistent. Yet human judgement is deeply affected by fatigue, stress, workload, and unconscious bias. Southampton's moderation policy refers to periods where staff are unable to work at normal capacity due to illness or pressure (University of Southampton, 2020). Under such conditions, consistency becomes uncertain rather than guaranteed.

Human grading is also shaped by implicit bias and subjective interpretation. Markers may unconsciously favour certain writing styles, cultural norms, or argument structures while penalising unfamiliar approaches. These influences are rarely visible or systematically audited. Unlike algorithms, human reasoning cannot easily be examined line by line to identify patterns of bias or inconsistency. Moderation procedures exist precisely because universities recognise that individual judgement alone is unreliable.

Sartor and Lagioia (2020) make an important point here: even a biased algorithmic system may be fairer than a more biased human decision maker because algorithmic bias can at least be measured, identified, and corrected. This does not mean algorithms are neutral. Roselli, Matthews and Talagala (2019) show that AI systems frequently reproduce existing social prejudices if trained on biased data. However, they also explain that algorithmic bias can be monitored through validation, statistical testing, and controlled evaluation.

Human bias is far more difficult to identify because it exists within personal experience, institutional culture, and unconscious assumptions. Markers often sincerely believe they are objective, which makes their biases less visible and less likely to be challenged. There is no equivalent of an audit log for human thought processes.

This does not mean AI grading is automatically superior. Poorly designed systems can reproduce harmful patterns at scale and create feedback loops that disadvantage students. However, the ethical comparison is not between perfect humans and flawed machines. It is between two imperfect systems that fail in different ways. AI may reduce some forms of inconsistency, but only if strong oversight and governance remain in place.

The ethical question therefore shifts from whether AI makes mistakes to whether its mistakes are more manageable, measurable, and correctable than those produced by existing human systems.

Reliability, Hallucination, and Accountability

A major criticism of AI grading is that LLMs are unreliable. These systems hallucinate, fabricate reasoning, and generate inaccurate feedback with apparent confidence. Hicks, Humphries and Slater (2024) argue that LLMs produce language based on patterns rather than truth, meaning their outputs may sound persuasive even when they are incorrect. Santos and Radanliev (2024) similarly note that AI systems often operate as opaque "black boxes" where even developers cannot clearly trace how a particular output was generated. Burrell (2016) describes this as a form of opacity that prevents ordinary human interpretation.

In grading, this creates serious risks. An AI system may misunderstand a student's argument, reward superficial features, or produce misleading feedback without anyone being able to explain why. If students cannot understand or challenge these decisions, trust in assessment weakens significantly.

Yet unreliability is not unique to AI. Human markers also misread essays, overlook arguments, and apply standards inconsistently. Thaler and Sunstein (2008) demonstrate how human decision making is shaped by cognitive biases and mental shortcuts that people are often unaware of. Sartor and Lagioia (2020) similarly note that human decision makers carry their own forms of prejudice and inconsistency.

The central ethical issue is therefore not imperfection itself but unaccountable imperfection. Decisions become ethically problematic when they cannot be explained, challenged, or corrected. Sparrow and Flenady (2025) argue that machines cannot bear moral responsibility for grading decisions. Responsibility must remain with human educators who can explain and defend outcomes. Kudina and van de Poel (2024) describe AI as part of a wider sociotechnical system where outcomes depend on interactions between humans, institutions, and technology rather than on the machine alone.

For this reason, AI should only play a supporting role. It may help identify inconsistencies, draft preliminary evaluations, or assist with administrative tasks, but final grading authority must remain with human markers. Students should always be able to challenge grades and receive explanations from a person. Santos and Radanliev (2024) support this hybrid governance model where AI systems remain supervised rather than autonomous.

Transparency, Explainability, and Justice

If AI is involved in grading, students must be able to understand how decisions are made. Hayes, van de Poel and Steen (2023) distinguish between transparency of the system, meaning how the tool works technically, and transparency concerning the system, meaning how it is governed and monitored. Both forms are essential in educational assessment.

A grading process that students cannot meaningfully question creates procedural injustice even if outcomes appear accurate. Sartor and Lagioia (2020), discussing GDPR and automated decision making, argue that individuals affected by algorithmic decisions should receive meaningful explanations and opportunities to contest outcomes. Applied to universities, this means students must be able to ask why they received a certain grade and receive a clear answer.

However, human grading is often far from transparent as well. Feedback can be vague, inconsistent, or shaped by tacit expectations that markers themselves struggle to explain. Southampton's assessment descriptors acknowledge that students do not fit into "neat boxes" (University of Southampton, n.d.-a), recognising the limits of rigid rubrics. Yet this flexibility can also leave students uncertain about why they received a particular mark.

Some transparency concerns may become more manageable with properly governed AI systems. Santos and Radanliev (2024) propose Artificial Intelligence Bills of Materials (AI BOMs), which would document how systems were built, trained, and evaluated. Such documentation could improve traceability and create stronger foundations for accountability. If systems are reviewable and auditable, decisions become easier to challenge and correct.

The debate therefore becomes less about rejecting technology entirely and more about institutional governance and responsibility.

Stakeholder Responsibility and Institutional Motives

A deeper issue behind AI grading concerns why universities want to adopt it in the first place. Public justifications usually focus on fairness and consistency, but financial and managerial pressures also play an important role. Universities face demands to reduce costs, manage growing student numbers, and process assessments more efficiently. Santos and Radanliev (2024) note that economic incentives often favour automation because it reduces labour costs and increases scalability. Sartor and Lagioia (2020) similarly argue that automation tends to reshape work in ways that gradually reduce human involvement.

If the primary motivation behind AI grading is efficiency rather than educational quality, the ethical concerns become more serious. Assessment risks being shaped by institutional convenience instead of student learning.

There are also concerns around surveillance and data use. AI grading systems may collect large amounts of information about students, including writing patterns, feedback histories, and behavioural data. Sartor and Lagioia (2020) warn that automated evaluation systems can create forms of persistent monitoring and behavioural profiling. Miller (2022) describes students in such systems as passive stakeholders because they are heavily affected by decisions while having little influence over how the systems are designed or used.

Still, these risks are not inevitable consequences of AI itself. Kudina and van de Poel (2024) argue that ethical outcomes depend heavily on the institutional and governance structures surrounding technology. Problems often emerge from poor oversight, weak accountability, or misaligned incentives rather than from the existence of algorithms alone.

Responsibility therefore belongs to multiple actors: universities, developers, regulators, and academic staff. Miller (2022) warns that when responsibility becomes too widely distributed, accountability weakens because each actor assumes someone else is responsible. This diffusion of responsibility may itself be one of the greatest ethical dangers.

Toward Cognitive Augmentation, Not Replacement

If the core issue is governance rather than technology itself, the most ethical approach may be neither full automation nor complete rejection. Nyholm (2024) describes a model of intellectual augmentation where AI supports human cognitive work rather than replacing it entirely.

In practice, this could involve AI assisting with repetitive or administrative tasks such as checking rubric coverage, identifying inconsistencies across large groups of assessments, or flagging contradictory feedback. Sartor and Lagioia (2020) note that algorithmic tools may reduce certain human psychological biases, while Hayes, van de Poel and Steen (2023) argue that carefully designed systems can make grading patterns more visible to moderators.

Importantly, this does not require handing final authority to machines. Instead, AI could reduce administrative burden and allow educators to focus more on genuinely human aspects of education: mentoring students, engaging critically with arguments, and providing meaningful feedback tailored to individuals.

However, this approach only remains ethical if AI outputs are treated as provisional rather than authoritative. Durán and Jongsma (2021) warn against giving excessive epistemic trust to opaque systems whose reasoning cannot be fully understood. Santos and Radanliev (2024) similarly stress that humans must retain interpretive authority over final outcomes. The moment institutions begin treating AI generated grades as unquestionable, the safeguards collapse.

Conclusion

Concerns about AI grading are entirely justified. Hallucination, opacity, weak accountability, and the possible removal of human judgement from assessment all create serious ethical risks. These concerns should not be dismissed.

At the same time, the analysis shows that human grading is not the perfectly fair alternative it is often assumed to be. Human markers are influenced by fatigue, workload pressure, inconsistency, and implicit bias. The real comparison is therefore not between flawless humans and flawed machines, but between two imperfect systems with different weaknesses.

AI grading only becomes ethically defensible under strict conditions. Human educators must remain accountable for final decisions. Systems must be transparent, reviewable, and open to challenge. Fairness must be continuously monitored rather than assumed. Most importantly, students must retain the ability to question outcomes and receive explanations from people rather than algorithms.

The deeper ethical issue is therefore not whether AI makes mistakes, but whether universities are willing to govern these systems responsibly and transparently. Ethical assessment is unlikely to be fully human or fully automated. It is more likely to depend on a carefully managed partnership between human judgement and artificial intelligence, where neither side operates without meaningful oversight.

Reference List

Burrell, J. (2016) ‘How the machine “thinks”: Understanding opacity in machine learning algorithms’, Big Data & Society, 3(1), pp. 1–12. doi: 10.1177/2053951715622512.

Durán, J.M. and Jongsma, K.R. (2021) ‘Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI’, Journal of Medical Ethics, 47(5), pp. 329–335. doi: 10.1136/medethics-2020-106820.

Hayes, P., van de Poel, I. and Steen, M. (2023) ‘Moral transparency of and concerning algorithmic tools’, AI and Ethics, 3(2), pp. 585–600. doi: 10.1007/s43681-022-00190-4.

Hicks, M.T., Humphries, J. and Slater, J. (2024) ‘ChatGPT is bullshit’, Ethics and Information Technology, 26(2), art. 38. doi: 10.1007/s10676-024-09775-5.

Kudina, O. and van de Poel, I. (2024) ‘A sociotechnical system perspective on AI’, Minds and Machines, 34(2), art. 21. doi: 10.1007/s11023-024-09680-2.

Miller, G.J. (2022) ‘Stakeholder roles in artificial intelligence projects’, Project Leadership and Society, 3, art. 100068. doi: 10.1016/j.plas.2022.100068.

Nyholm, S. (2024) ‘Artificial Intelligence and Human Enhancement: Can AI Technologies Make Us More (Artificially) Intelligent?’, Cambridge Quarterly of Healthcare Ethics, 33(1), pp. 76–88. doi: 10.1017/S0963180123000464.

Roselli, D., Matthews, J. and Talagala, N. (2019) ‘Managing Bias in AI’, Companion Proceedings of The 2019 World Wide Web Conference (WWW ’19 Companion), pp. 539–544. doi: 10.1145/3308560.3317590.

Santos, O. and Radanliev, P. (2024) Beyond the Algorithm: AI, Security, Privacy, and Ethics. Upper Saddle River: Addison-Wesley Professional. Available at: O’Reilly Learning Platform (Accessed: 15 May 2026).

Sartor, G. and Lagioia, F. (2020) The impact of the General Data Protection Regulation (GDPR) on artificial intelligence. European Parliamentary Research Service (EPRS). Available at: European Parliament Think Tank (Accessed: 15 May 2026).

Sparrow, R. and Flenady, G. (2025) ‘The Testimony Gap: Machines and Reasons’, Minds and Machines, 35(1), art. 12. doi: 10.1007/s11023-025-09712-5.

Thaler, R.H. and Sunstein, C.R. (2008) Nudge: Improving Decisions about Health, Wealth, and Happiness. New Haven: Yale University Press.

University of Southampton (2020) Double-Blind Marking and Moderation Policy (COVID19). Available at: University of Southampton (Accessed: 15 May 2026).

University of Southampton (2023) Assessment Feedback to Students Policy. Available at: University of Southampton (Accessed: 15 May 2026).

University of Southampton (n.d.-a) Assessment Descriptors. Quality Handbook. Available at: University of Southampton Assessment Descriptors (Accessed: 15 May 2026).

University of Southampton (n.d.-b) Assessment. Quality Handbook. Available at: University of Southampton Assessment Framework (Accessed: 15 May 2026).

Appendices

Appendix A: Introduction Evidence

A1 University of Southampton Academic Judgement Policy
(University of Southampton, n.d.-b)

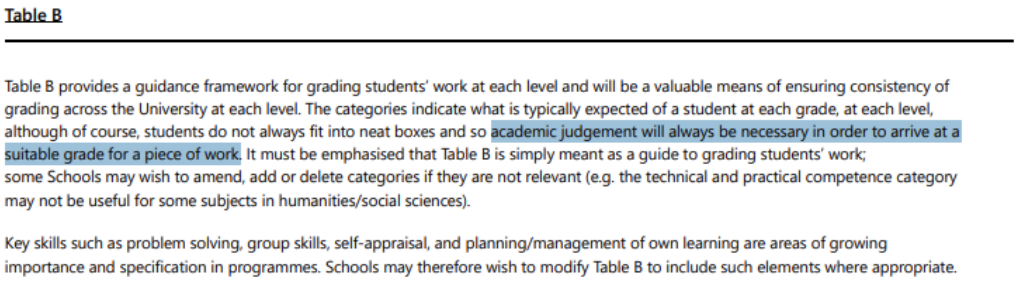
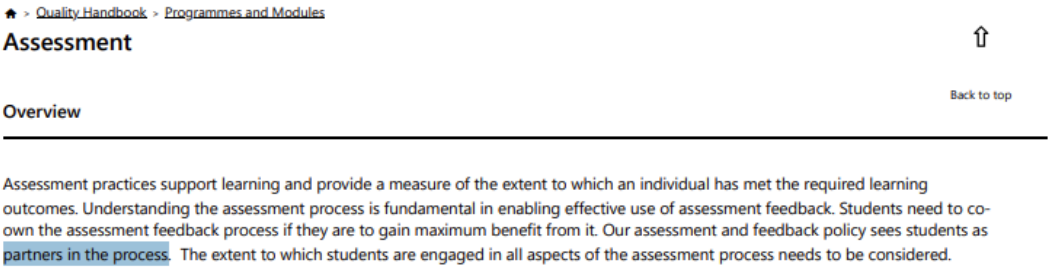


Table B

Table B provides a guidance framework for grading students' work at each level and will be a valuable means of ensuring consistency of grading across the University at each level. The categories indicate what is typically expected of a student at each grade, at each level, although of course, students do not always fit into neat boxes and so academic judgement will always be necessary in order to arrive at a suitable grade for a piece of work. It must be emphasised that Table B is simply meant as a guide to grading students' work; some Schools may wish to amend, add or delete categories if they are not relevant (e.g. the technical and practical competence category may not be useful for some subjects in humanities/social sciences).

Key skills such as problem solving, group skills, self-appraisal, and planning/management of own learning are areas of growing importance and specification in programmes. Schools may therefore wish to modify Table B to include such elements where appropriate.

A2 Southampton Feedback Policy
(University of Southampton, 2023)



Quality Handbook > Programmes and Modules

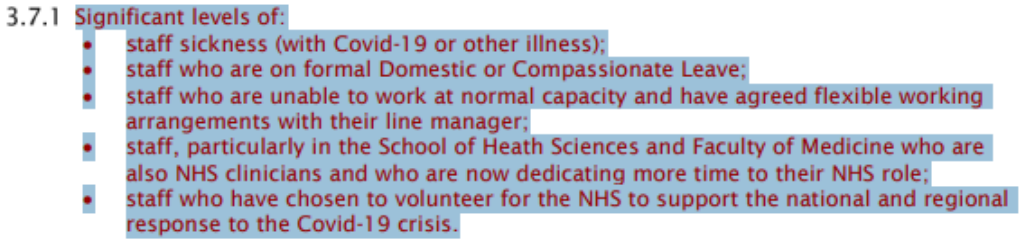
Assessment

Back to top

Overview

Assessment practices support learning and provide a measure of the extent to which an individual has met the required learning outcomes. Understanding the assessment process is fundamental in enabling effective use of assessment feedback. Students need to co-own the assessment feedback process if they are to gain maximum benefit from it. Our assessment and feedback policy sees students as partners in the process. The extent to which students are engaged in all aspects of the assessment process needs to be considered.

A3 Southampton Moderation / Staffing Policy
(University of Southampton, 2020)
Used to infer workload strain affecting grading consistency.



3.7.1 Significant levels of:

- staff sickness (with Covid-19 or other illness);
- staff who are on formal Domestic or Compassionate Leave;
- staff who are unable to work at normal capacity and have agreed flexible working arrangements with their line manager;
- staff, particularly in the School of Heath Sciences and Faculty of Medicine who are also NHS clinicians and who are now dedicating more time to their NHS role;
- staff who have chosen to volunteer for the NHS to support the national and regional response to the Covid-19 crisis.

A4 Sartor and Lagioia (2020)

offs among competing values. Algorithms are not only a threat to be regulated; with the right safeguards in place, they have the potential to be a positive force for equity.[37]

In conclusion, it seems that issues that have just been presented should not lead us to exclude categorically the use of automated decision-making. The alternative to automated decision-making is not perfect decisions but human decisions with all their flaws: a biased algorithmic system can still be fairer than an even more biased human decision-maker. In many cases, the best solution consists in integrating human and automated judgements, by enabling the affected individuals to request a human review of an automated decision as well as by favouring transparency and developing methods and technologies that enable human experts to analyse and review automated decision-making. In fact, AI systems have demonstrated an ability to successfully also act in domains traditionally entrusted the trained intuition and analysis of humans, such as medical diagnosis, financial investment, the granting of loans, etc. The future challenge will consist in finding the best combination between human and automated intelligence, taking into account the capacities and the limitations of both.

A5 Hicks, Humphries and Slater (2024)

Abstract

Recently, there has been considerable interest in large language models: machine learning systems which produce human-like text and dialogue. Applications of these systems have been plagued by persistent inaccuracies in their output; these are often called "AI hallucinations". We argue that these falsehoods, and the overall activity of large language models, is better understood as *bullshit* in the sense explored by Frankfurt (On Bullshit, Princeton, 2005): the models are in an important way indifferent to the truth of their outputs. We distinguish two ways in which the models can be said to be bullshitters, and argue that they clearly meet at least one of these definitions. We further argue that describing AI misrepresentations as bullshit is both a more useful and more accurate way of predicting and discussing the behaviour of these systems.

A6 Santos and Radanliev (2024)

Lack of Transparency	Many AI and ML models are complex and are frequently referred to as "black boxes" because it can be difficult to understand how they make decisions or forecast future events. Accountability issues result from this lack of openness mainly because it is more challenging to identify and correct mistakes or prejudices.

To protect moral judgment and avoid an overreliance on computers, it is essential to maintain human monitoring and control over AI systems. In this part, we

[redacted]

go into the methods that allow for human oversight and management of AI operations. We investigate methods to enable humans to maintain complete control over AI systems, ranging from interpretability and explainability approaches that give information on AI decision-making processes to developing distinct limits and red lines. In addition, we emphasize the significance of continuing governance mechanisms that minimize possible risks and assure ethical AI use.

A7 Miller (2022)

A B S T R A C T

Algorithmic decision-making implemented through artificial intelligence (AI) projects is augmenting or replacing human decision-making across numerous industries. Although AI systems may impact individuals and society in life-and-death situations, project organizations or plans may ignore the concerns of the passive stakeholders. This research describes the components, lifecycle, and characteristics of AI projects. It employs stakeholder theory and a systematic literature review with thematic analysis to identify and classify individuals, groups, and organizations into six stakeholder project roles. It configures the stakeholder salience model with a harm attribute to identify passive stakeholders—individuals affected by AI systems but powerless to affect the project—and their nexus to the project. The study contributes a novel method for identifying passive stakeholders and highlights the need to engage developers, operators, and representatives of passive stakeholders to achieve moral, ethical, and sustainable development.

Appendix B: Why Human-Only Grading Appears Ethically Preferable

B1 Southampton Assessment Framework

Table B

Table B provides a guidance framework for grading students' work at each level and will be a valuable means of ensuring consistency of grading across the University at each level. The categories indicate what is typically expected of a student at each grade, at each level, although of course, students do not always fit into neat boxes and so academic judgement will always be necessary in order to arrive at a suitable grade for a piece of work. It must be emphasised that Table B is simply meant as a guide to grading students' work; some Schools may wish to amend, add or delete categories if they are not relevant (e.g. the technical and practical competence category may not be useful for some subjects in humanities/social sciences).

B2 Southampton "Partners in the Process" Policy

Assessment practices support learning and provide a measure of the extent to which an individual has met the required learning outcomes. Understanding the assessment process is fundamental in enabling effective use of assessment feedback. Students need to co-own the assessment feedback process if they are to gain maximum benefit from it. Our assessment and feedback policy sees students as partners in the process. The extent to which students are engaged in all aspects of the assessment process needs to be considered.

B3 Sparrow and Flenady (2025)

Abstract

Most people who have considered the matter have concluded that machines cannot be moral agents. Responsibility for acting on the outputs of machines must always rest with a human being. A key problem for the ethical use of AI, then, is to ensure that it does not block the attribution of responsibility to humans or lead to individuals being unfairly held responsible for things over which they had no control. This is the "responsibility gap". In this paper, we argue that the claim that machines cannot be held responsible for their actions has unacknowledged implications for the conditions under which the outputs of AI can serve as reasons for belief. Following Robert Brandom, we argue that, because the assertion of a claim

B4 Burrell (2016)

writing (and reading) code is a specialist skill and; (3) an opacity that stems from the mismatch between mathematical optimization in high-dimensionality characteristic of machine learning and the demands of human-scale reasoning and styles of semantic interpretation. This third form of opacity (often conflated with the

semantic explanations. The examples of handwriting recognition and spam filtering helped to illustrate how the workings of machine learning algorithms can escape full understanding and interpretation by humans, even for those with specialized training, even for computer scientists.

B5 Hicks, Humphries and Slater (2024)

Abstract

Recently, there has been considerable interest in large language models: machine learning systems which produce human-like text and dialogue. Applications of these systems have been plagued by persistent inaccuracies in their output; these are often called "AI hallucinations". We argue that these falsehoods, and the overall activity of large language models, is better understood as *bullshit* in the sense explored by Frankfurt (On Bullshit, Princeton, 2005): the models are in an important way indifferent to the truth of their outputs. We distinguish two ways in which the models can be said to be bullshitters, and argue that they clearly meet at least one of these definitions. We further argue that describing AI misrepresentations as bullshit is both a more useful and more accurate way of predicting and discussing the behaviour of these systems.

their training data. No reasoning is involved […] Similarly, language models are prone to making stuff up […] because they are not designed to express some underlying set of information in natural language; they are only manipulating the form of language" (Shah & Bender, 2022). These models aren't designed to transmit information, so we shouldn't be too surprised when their assertions turn out to be false.

B6 Santos and Radanliev (2024)

Overreliance

Overreliance on AI and LLMs occurs when individuals or systems lean too heavily on these models for decision-making or content creation, often sidelining critical oversight. LLMs, while remarkable in generating imaginative and insightful content, are not infallible. They can, at times, produce outputs that are inaccurate, unsuitable, or even harmful. Such instances, known as *hallucinations* or *confabulations*, have the potential to spread false information, lead to misunderstandings, instigate legal complications, and tarnish reputations.

Appendix C: Questioning the Assumption of Human Objectivity

C1 Southampton Moderation Policy

3.7.1 Significant levels of:

- staff sickness (with Covid-19 or other illness);
- staff who are on formal Domestic or Compassionate Leave;
- staff who are unable to work at normal capacity and have agreed flexible working arrangements with their line manager;
- staff, particularly in the School of Heath Sciences and Faculty of Medicine who are also NHS clinicians and who are now dedicating more time to their NHS role;
- staff who have chosen to volunteer for the NHS to support the national and regional response to the Covid-19 crisis.

For examination scripts, double-blind marking is not expected though moderation is required (see section on Moderation below).

C2 Sartor and Lagioia (2020)

In conclusion, it seems that issues that have just been presented should not lead us to exclude categorically the use of automated decision-making. The alternative to automated decision-making is not perfect decisions but human decisions with all their flaws: a biased algorithmic system can still be fairer than an even more biased human decision-maker. In many cases, the best solution consists in integrating human and automated judgements, by enabling the affected individuals to request a

[W]ith appropriate requirements in place, the use of algorithms will make it possible to more easily examine and interrogate the entire decision process, thereby making it far easier to know whether discrimination has occurred. By forcing a new level of specificity, the use of algorithms also highlights, and makes transparent, central trade-

Some scholars have observed that in many domains automated predictions and decisions are not only cheaper, but also more precise and impartial than human ones. AI systems can avoid the typical fallacies of human psychology (overconfidence, loss aversion, anchoring, confirmation bias, representativeness heuristics, etc.), and the widespread human inability to process statistical data,[33] as well as typical human prejudice (concerning, e.g., ethnicity, gender, or social background). In many assessments and decisions – on investments, recruitment, creditworthiness, or also on judicial matters, such as bail, parole, and recidivism – algorithmic systems have often performed better, according to usual standards, than human experts.[34]

C3 Roselli, Matthews and Talagala (2019)

3. Managing Bias

Given the number of potential sources of bias currently known (and more are being discovered as the field matures), it can be daunting to face how to tackle them. Considering the myriad of issues that cause bias to enter the system, we cannot expect a single approach to resolve all of them. Instead, we propose a combination of quantitative assessments, business processes, monitoring, data review, evaluations, and controlled experiments. Before detailing the above stages, we establish some ground rules for processes we want to include.

First, any process to evaluate an AI system for bias must be achievable by those that are not necessarily the primary developer. This is important because the system may need to be understood by non-technical management or evaluated by an auditor or regulator. In fact, since complex models are increasingly available from external sources [16, 17], even the primary developer may not know the model's inner workings. Therefore, we require that our processes not require any knowledge of the internal workings of the model itself; we treat the entire AI pipeline (including any feature engineering) as a black box and use only the input and output data for evaluation.

Second, transparency of input data is important. This is the only way to verify that the data is accurate and does not contain protected or incorrect data. Even if the data is private, it should be viewable by the person about whom the data is [18]. Methodologies for data versioning, data cataloging, and data tracking and governance are also critical to ensure that the specific da-

input features determined the resulting prediction [15, 22]. Such predictions can be validated by providing a qualified judge (who does not a priori know the AI system's prediction) with the same feature values and comparing the judge's determination with the AI's prediction. This both validates and provides explanations for predictions. Such an evaluation process can help identify both non-generalizable features and irrelevant correlations.

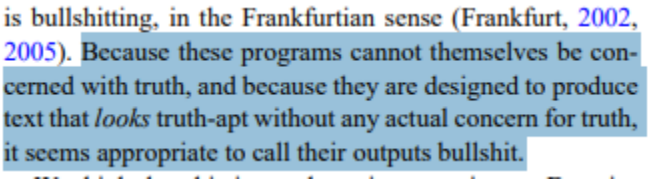
3.3.3. Match Training to Production Needs. It is important to monitor the distribution of input data to see whether production data is consistent with the data used in training. Input data should be monitored for anomalous cases that differ substantially from those seen in training. Input streams should also be monitored to ensure that the distribution of data seen in production does not stray from the distribution anticipated by the training data set.

3.6. Quantify Feedback Loops

In cases for which the AI predictions can impact its own future input, it is important to ensure that this impact is quantified by comparing against a suitable control or comparing the results against external evidence. For example, algorithms that predict crime based on police reports may cause more police to be deployed to the site of the predictions, which in turn may result in more police reports [7]. Note that not all feedback loops are considered negative; algorithms that return search results may return better search results when less common search terms are used, which may result in users learning to enter less common search terms. However, even in positive cases, it is important for the designers to understand the impact of feedback loops.

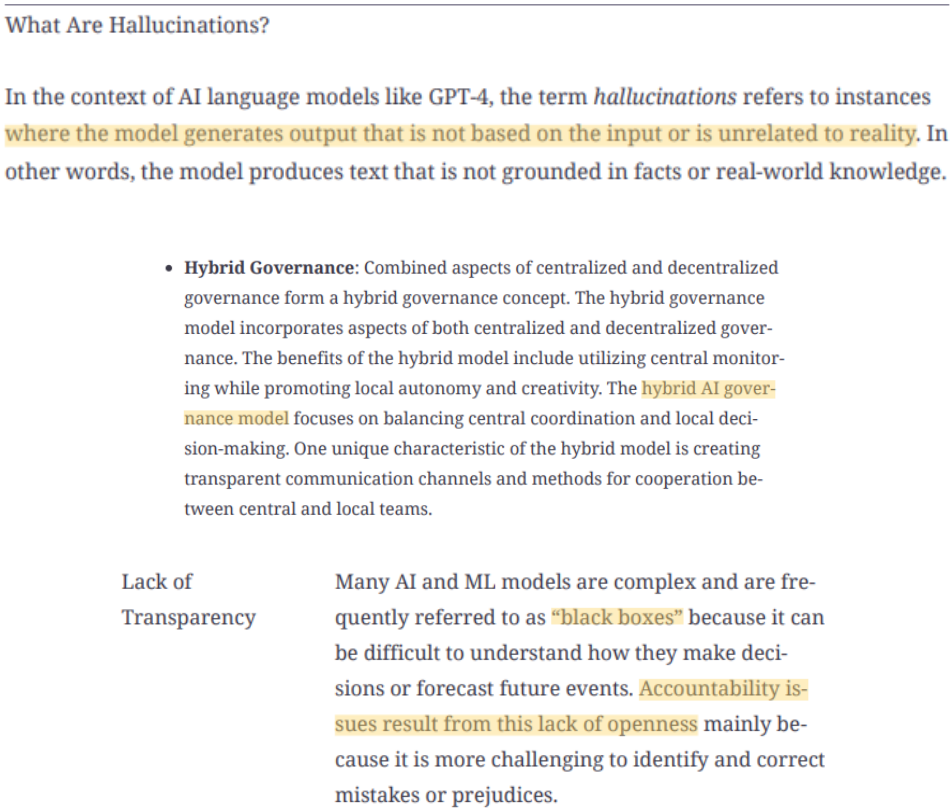
Appendix D: Reliability, Hallucination, and Accountability

D1 Hicks, Humphries and Slater (2024)



is bullshitting, in the Frankfurtian sense (Frankfurt, 2002, 2005). Because these programs cannot themselves be concerned with truth, and because they are designed to produce text that *looks* truth-apt without any actual concern for truth, it seems appropriate to call their outputs bullshit.

D2 Santos and Radanliev (2024)



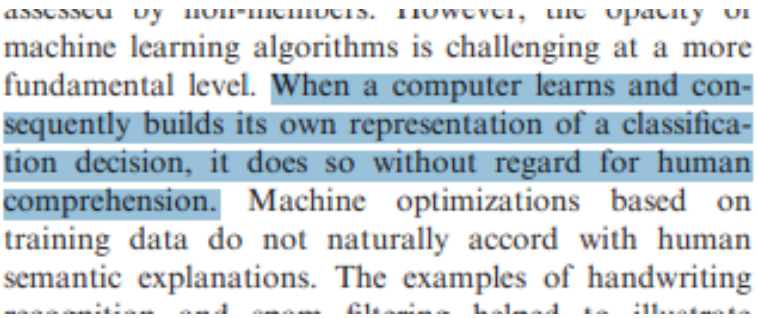
What Are Hallucinations?

In the context of AI language models like GPT-4, the term *hallucinations* refers to instances where the model generates output that is not based on the input or is unrelated to reality. In other words, the model produces text that is not grounded in facts or real-world knowledge.

- **Hybrid Governance**: Combined aspects of centralized and decentralized governance form a hybrid governance concept. The hybrid governance model incorporates aspects of both centralized and decentralized governance. The benefits of the hybrid model include utilizing central monitoring while promoting local autonomy and creativity. The hybrid AI governance model focuses on balancing central coordination and local decision-making. One unique characteristic of the hybrid model is creating transparent communication channels and methods for cooperation between central and local teams.

Lack of Transparency	Many AI and ML models are complex and are frequently referred to as "black boxes" because it can be difficult to understand how they make decisions or forecast future events. Accountability issues result from this lack of openness mainly because it is more challenging to identify and correct mistakes or prejudices.

D3 Burrell (2016)



assessed by non-members. However, the opacity of machine learning algorithms is challenging at a more fundamental level. When a computer learns and consequently builds its own representation of a classification decision, it does so without regard for human comprehension. Machine optimizations based on training data do not naturally accord with human semantic explanations. The examples of handwriting

writing (and reading) code is a specialist skill and; (3) an opacity that stems from the mismatch between mathematical optimization in high-dimensionality characteristic of machine learning and the demands of human-scale reasoning and styles of semantic interpretation. This third form of opacity (often conflated with the second form as part of the general sense that algorithms

D4 Thaler and Sunstein (2008)

thing, choosing a retirement fund or choosing among medical treatment options represent quite more challenging choices. When faced with challenging choices, on result is that people often make choices that are not in their best interest. They make mistakes.

D5 Sartor and Lagioia (2020)

have made under the same circumstances. In the second case, the system will predict how a certain choice would affect the goals being pursued (preventing non-payments, preventing recidivism). In the first case the system would reproduce the virtues – accuracy, impartiality, farness – but also the

11

vices – carelessness, partiality, unfairness – of the humans it is imitating. In the second case it would more objectively approximate the intended outcomes.

In conclusion, it seems that issues that have just been presented should not lead us to exclude categorically the use of automated decision-making. The alternative to automated decision-making is not perfect decisions but human decisions with all their flaws: a biased algorithmic system can still be fairer than an even more biased human decision-maker. In many cases, the best solution consists in integrating human and automated judgements, by enabling the affected individuals to request a human review of an automated decision as well as by favouring transparency and developing methods and technologies that enable human experts to analyse and review automated decision-making. In fact, AI systems have demonstrated an ability to successfully also act in domains traditionally entrusted the trained intuition and analysis of humans, such as medical diagnosis, financial investment, the granting of loans, etc. The future challenge will consist in finding the best combination between human and automated intelligence, taking into account the capacities and the limitations of both.

D6 Sparrow and Flenady (2025)

As we noted above, it is an article of faith in most discussions of AI ethics that machines cannot be moral agents: responsibility for the consequences of the "decisions" of AI always comes back to one or more human beings. Here, for instance, are some representative discussions of the matter, which, although they disagree on the reasons why, concur that machines cannot be moral agents:

The main reason, we believe, that the existence of the testimony gap has gone unrecognised is that in almost every circumstance in which we encounter, and (apparently) rely on machines, another human being has already "vouched for" the machine. That is, a human being must assume responsibility – or be held responsible – *for* the epistemic quality of the machine's output. We think that we are relying on the machine,

D7 Kudina and van de Poel (2024)

In their paper "Contestable AI by design: towards a framework," Alfrink and colleagues put contestability as a central feature of the interwoven character of AI systems. Understood as a design feature that underlies the interaction of the social, technological and institutional components of an AI technology, contestability here assumes a system's responsiveness to user's challenge of the system and embeds human rights to self-determination and control in the system. Relying on a mix of methods, such as systematic literature review, the authors propose a framework that would combine essential features (e.g. human review and intervention requests) and practices (e.g. agonistic approaches to ML development) in the AI system to make it contestable. While this framework does not pretend to be exhaustive, it points to the challenges of the decentralized sociotechnical AI systems and at the same time, offers opportunities for system improvement by relying on conflicting points in AI development, leveraging dissent as a rich resource.

Appendix E: Transparency, Explainability, and Justice

E1 Hayes, van de Poel and Steen (2023)

importance of transparency, that is, how it can cast light on the ethical acceptability of X and how such knowledge can empower human agency and verify or highlight value supports or conflicts and allow us to respond to them. Therefore, more specifically, we will argue that:

y

policy. Here, we attempt to further clarify transparency. We argue that transparency is the state of affairs that obtains when relevant and understandable information about some X is available and accessible to some target audience (A), so that this information is sufficient for A for the purpose (P). Moreover, we connect this conceptualisation with transparency's moral value, where P is to provide an account about X's supportive or conflicting relationship with relevant values and goals. Such teleological ends in our context here can be the ability to account for the degree to which an algorithm, process or organisation respects certain values and is conducive to (social) goals.

E2 Sartor and Lagioia (2020)

- Guidance is needed on profiling and automated decision-making. It seems that an obligation of reasonableness – including normative and reliability aspects – should be imposed on controllers engaging in profiling, mostly, but not only when profiling is aimed at automated decision-making. Controllers should also be under an obligation to provide individual explanations, to the extent that this is possible according to the available AI technologies, and reasonable according to costs and benefits. The explanations may be high-level, but they should still enable users to contest detrimental outcomes.

- Fairness: it should be intended under its substantive and procedural dimension. The substantive dimension implies a commitment to: ensuring equal and just distribution of

---

both benefits and costs, and ensuring that individuals and groups are free from unfair bias, discrimination and stigmatisation. The procedural dimension entails the ability to contest and seek effective redress against decisions made by AI systems and by the humans operating them.

E3 Kudina and van de Poel (2024)

In their paper "Contestable AI by design: towards a framework," Alfrink and colleagues put contestability as a central feature of the interwoven character of AI systems. Understood as a design feature that underlies the interaction of the social, technological and institutional components of an AI technology, contestability here assumes a system's responsiveness to user's challenge of the system and embeds human rights to self-determination and control in the system. Relying on a mix of methods, such as systematic literature review, the authors propose a framework that would combine essential features (e.g. human review and intervention requests) and practices (e.g. agonistic approaches to ML development) in the AI system to make it contestable. While this framework does not pretend to be exhaustive, it points to the challenges of the decentralized sociotechnical AI systems and at the same time, offers opportunities for system improvement by relying on conflicting points in AI development, leveraging dissent as a rich resource.

E4 Santos and Radanliev (2024)

AI bill of materials (AI BOMs) provide a comprehensive inventory of all the components, data, algorithms, and tools that are used in building and deploying an AI system. Just as traditional manufacturing relies on BOMs to detail parts, specifications, and sources for products, AI BOMs ensure transparency, traceability, and accountability in AI development and its supply chain. By documenting every element in an AI solution, from the data sources used for training to the software libraries integrated, AI BOMs enable developers, auditors, and stakeholders to assess the quality, reliability, and security of the system. Furthermore, in cases of system failures, biases, or security breaches, AI BOMs can facilitate swift identification of the problematic component, promoting responsible AI practices and maintaining trust among users and the industry.

- **Hybrid Governance**: Combined aspects of centralized and decentralized governance form a hybrid governance concept. The hybrid governance model incorporates aspects of both centralized and decentralized governance. The benefits of the hybrid model include utilizing central monitoring while promoting local autonomy and creativity. The hybrid AI governance model focuses on balancing central coordination and local decision-making. One unique characteristic of the hybrid model is creating transparent communication channels and methods for cooperation between central and local teams.

Appendix F: Stakeholder Responsibility and Institutional Motives

F1 Santos and Radanliev (2024)

Unemployment and Job Displacement	The automation potential of AI and ML may result in significant changes to the workforce, including job losses. The effect on people's livelihoods and the obligation to offer assistance and retraining chances to those impacted create ethical questions.

F2 Sartor and Lagioia (2020)

However, the development of AI and its convergence with big data also leads to serious risks for individuals, for groups, and for the whole of society. For one thing, AI can eliminate or devalue the jobs of those who can be replaced by machines: many risk losing the 'race against the machine',[29] and therefore being excluded from or marginalised in the job market. This may lead to poverty and social exclusion, unless appropriate remedies are introduced (consider, for instance, the future impact of autonomous vehicles on taxi and truck drivers, or the impact of smart chatbots on call-centres workers).

Moreover, by enabling big tech companies to make huge profits with a limited workforce, AI contributes to concentrating wealth in those who invest in such companies or provide them with high-level expertise. This trend favours economic models in which *'the winner takes all'*. Within

concerned.[45] According to Zuboff, we have not yet developed adequate legal, political or social measures by which to check the potentially disruptive outcomes of surveillance capitalism and keep them in balance. However, she observes, the GDPR could be an important step in this direction, as a 'springboard to challenging the legitimacy of surveillance capitalism and ultimately vanquishing its instrumentarian power', towards 'society's rejection of markets based on the dispossession of human experience as a means to the prediction and control of human behavior for others' profit.'

F3 Miller (2022)

- *Harm* evaluates if the stakeholder would be harmed by the implementation or use of the system. Here we can build from ethical issues identified by Someh et al. (2019) for big data analytics for individuals, organizations, and society and classify harms, damages, or losses as follows: harms: bodily harm, loss of life, limitation of rights (freedom of movement), surveillance, psychological; loss: violations of human or civil rights such as loss of privacy, security, freedom, financial, job; damage: trust, reputation, environment.

code. The development process can create a moral buffer where no one is accountable for the impacts of system usage on individuals and society (Moser et al., 2022; Singh et al., 2019). A moral buffer occurs when neither the team who developed the algorithms nor the human decision-makers who use the algorithms take responsibility for the social impact (Green and Chen, 2019). Thus, this study provides further

Artificial intelligence (AI) adoption is growing rapidly, and AI is expected to revolutionize and transform production processes, thereby influencing the behavior of economic actors (OECD, 2019). Scholars have argued that system developers are moral agents responsible for the impacts of their systems on individuals and society (Cohen et al., 2014; Manders-Huits, 2006; Martin, 2019; Miller, 2022; Wieringa, 2020). A

F4 Kudina and van de Poel (2024)

A third element is governance. There is now a lot of attention for AI ethics, and rightly so, but properly dealing with the ethical and social issues raised by AI systems is not just a matter of ethical guidelines and proper design, it also requires attention for broader governance and political issues. Given the complexity of sociotechnical systems, dealing with the disruptive potential of AI requires more than just ethics, it requires a range of technical, social, economic and governance choices that are coordinated; in other words, it requires politics.

Appendix G: Toward Cognitive Augmentation, Not Replacement

G1 Nyholm (2024)

Specifically, the idea that AI-driven recommender systems could function as a form of moral enhancers has been addressed by authors such as Alberto Giubilini, Julian Savulescu, and Michał Klincewicz.[4] Relatedly, Alexandre Erler and Vincent Müller have argued that AI can function as an augmentation of human intelligence.[5] My aim here is to add to these deliberations by relating the perspective on AI as a form of enhancement to the discussion of AI as a form of cognitive extension. Additionally, the existing discussion is expanded by suggesting that AI might under certain circumstances give humans a form of artificial intelligence.

G2 Sartor and Lagioia (2020)

valuable commodities. In particular, AI enables automated decision-making even in domains that require complex choices, based on multiple factors and non-predefined criteria. In many cases, automated predictions and decisions are not only cheaper, but also more precise and impartial than human ones, as AI systems can avoid the typical fallacies of human psychology and can be subject to rigorous controls. However, algorithmic decisions may also be mistaken or discriminatory, reproducing human biases and introducing new ones. Even when automated assessments of

Thanks to AI, it may be possible to achieve a new cooperation between humans and machines, which overcomes the classical model in which machines only performed routine and repetitive tasks. This integration was already predicted in the early 1960s' by JK Licklider, a scientist who played a key role in the development of the Internet. He argued that in the future, cooperation between human and computer would include creative activities, i.e., 'making decisions and controlling

[25] Kurzweil (2012).

[26] McAfee, and Brynjolfsson (2019).

18

complex situations without inflexible dependence on predetermined programs.[27] Today, it is indeed possible to integrate humans and machines in new ways that not only exploit synergies, but may also preserve and enhance human initiative and work satisfaction.[28]

G3 Hayes, van de Poel and Steen (2023)

The advantage of ML algorithms is that they can extract actionable insights from large amounts of data [7], and thus they can effectively automate large-scale cognitive workloads towards solving problems.

G4 Durán and Jongsma (2021)

AI system are trusted because the algorithm itself is reliable, these should not be followed blindly without further assessment. Instead, we must keep humans in the loop of decision making by algorithms.[40 42] Even if black box algorithms become more technologically advanced and able to somehow include contextual factors and patient preferences in the assessment, there could still be good reasons not to follow the algorithm's recommendation blindly (see footnote 1). It follows that it is unlikely and undesirable for algorithms to replace physicians altogether, as some scholars have argued in favour of.[45–47]

G5 Santos and Radanliev (2024)

and other healthcare professionals dictate. As a result, healthcare professionals may devote more time to caring for patients while spending less time and effort on manual paperwork.

First and foremost, AI should not replace human judgment but rather provide a tool for making educated decisions. AI technologies can increase human autonomy by enhancing our understanding of complex problems; however, we must maintain the capacity to assess AI-generated suggestions critically and use our discretion.

Appendix H: Recommendations and Concluding Position

H1 Sparrow and Flenady (2025)

a human being (Johnson, 2006; Véliz, 2021). Machines cannot be full moral agents. That is to say, machines can never be held morally responsible for the consequences of their outputs. A key problem for the ethical use of AI, then, is to ensure that the use of AI does not block the attribution of responsibility to human beings or lead to individuals being unfairly held responsible for things over which they had no control. This is the notorious "responsibility gap" (Matthias, 2004; Sparrow, 2007).

H2 Santos and Radanliev (2024)

nance. The benefits of the hybrid model include utilizing central monitoring while promoting local autonomy and creativity. The hybrid AI governance model focuses on balancing central coordination and local decision-making. One unique characteristic of the hybrid model is creating

Overreliance on AI and LLMs occurs when individuals or systems lean too heavily on these models for decision-making or content creation, often sidelining critical oversight. LLMs, while remarkable in generating imaginative and insightful content, are not infallible. They can, at

manufacturing relies on BOMs to detail parts, specifications, and sources for products, AI BOMs ensure transparency, traceability, and accountability in AI development and its supply chain. By documenting every element in an AI solution, from the data sources used for training to the software libraries integrated, AI BOMs enable developers, auditors, and stakeholders to assess the quality, reliability, and security of the system. Furthermore, in cases of system failures, biases, or security breaches, AI BOMs can facilitate swift identification of the problematic component, promoting responsible AI practices and maintaining trust among users and the industry.

H3 Sartor and Lagioia (2020)

According to the first one, the European legislator, by only including the request for specific explanation in the recitals and omitting it from the articles of the GDPR, intended to convey a double message: to exclude an enforceable legal obligation to provide individual explanations, while recommending that data controllers provide such explanations when convenient, according to their discretionary determinations. Following this interpretation, providing individualised explanation would only be a good practice, and not a legally enforceable requirement.

H4 Roselli, Matthews and Talagala (2019)

daunting to face how to tackle them. Considering the myriad of issues that cause bias to enter the system, we cannot expect a single approach to resolve all of them. Instead, we propose a combination of quantitative assessments, business processes, monitoring, data review, evaluations, and controlled experiments. Before detailing the above stages, we establish some

---

Appendix I: Conclusion Synthesis

Human grading is imperfect

This conclusion was developed primarily in the section questioning human objectivity. Evidence included Southampton moderation policies discussing workload pressures and sources examining human cognitive bias and inconsistency, particularly Sartor and Lagioia (2020) and Thaler and Sunstein (2008).

AI systems create accountability and opacity risks

This conclusion emerged from the sections on reliability and transparency. Burrell (2016), Hicks, Humphries and Slater (2024), and Santos and Radanliev (2024) were used to support concerns about hallucination, black-box reasoning, and weak explainability.

Accountability must remain human

The argument that final grading responsibility should remain with educators was developed using Sparrow and Flenady (2025), particularly their discussion of machines lacking moral agency and genuine justificatory reasoning.

Transparency and contestability are necessary

The discussion of procedural justice and explainability drew mainly from Hayes, van de Poel and Steen (2023) and Sartor and Lagioia (2020), especially regarding meaningful explanation and the ability to challenge automated decisions.

AI should assist rather than replace

The recommendation for augmentation rather than replacement was informed by Nyholm (2024) alongside governance arguments from Santos and Radanliev (2024).