A Multimodal Artificial Intelligence Framework Using Behavioural and Natural Language Analysis for Detecting Digital Risks Among Adolescents

Pooja Banerjee[1], Neeraj Kumar[2]
[1,2]*Suresh Gyan Vihar University*

***Abstract*—Digital environments are becoming more prevalent in which harmful content, cyberbullying, sexual exploitation, coercive interactions, misinformation, privacy threats and other risk elements of the online world may manifest themselves through a combination or pattern of behaviour and language, and these risks are increasing with age among adolescents. Current detection methods tend to target only one type of signal like text classification or activity on a platform and might lack the insight to detect signals from one platform that are meaningful only when combined with another. The present paper suggests a multimodal artificial intelligence approach for digital-risk detection in adolescents that integrates behaviour analytics and natural language processing capabilities, while maintaining privacy and human supervision. The proposed architecture comprises five major layers: Privacy-Aware Data Collection, Behavioural Feature Extraction, Natural Language Analysis, Uncertainty-Aware Multimodal Fusion, Risk Assessment and HITL Intervention. The framework does not establish a diagnostic/punitive surveillance system but rather constitutes an early warning system. It includes protection of privacy, with regards to consent, data minimisation, bias and fairness, explainability, false positive, age-appropriate design and escalation procedures. The paper reviews recent literature on Adolescent Digital risks, Computational Detection, and Multimodal artificial intelligence to pinpoint a research gap related to integrated analysis of behavioral and linguistic signals in context. The proposed framework offers a foundation for future validation via ethical controlled datasets and evaluation. The software's key features are its safety-focused architecture, which integrates with technical detection, uncertainty assessment, explainability, and proportional human review.**

***Index Terms*—Adolescents, AI, multimodal learning, natural language processing, behavioural analysis, digital risks, online safety.**

I. INTRODUCTION

Digital technologies are now integrated into and are an important part of adolescents' communication, learning, play and social interactions. The opportunities are also great, but the risks are potentially great and may not be easily identified at an early age. How cyberbullying, harassment, grooming, sexual exploitation, self-harm-related posts, fraud, coercive communication, misinformation, privacy violations, and manipulative interactions can happen and can create different types of digital footprints can vary across platforms. These environments are complex and, in most cases, risk is not visible as a single signal. Rather, it can arise from the relationship between what a young person says, the actions of an account, the time order of interactions, and the setting(s) of interactions.

Artificial intelligence provides tools to process high volumes of non-homogeneous data, and to find patterns that could not be manually identified. Natural language processing can detect language patterns in posts, comments and conversations; behaviours can be detected from temporal, relational and activity-based patterns. But, when applied to each stream separately, these treatments can lead to incomplete or inappropriate assessments. Repeated exposure to a seemingly innocuous message, unexpected shifts in interaction styles, the transfer of accounts, and other contextual factors can give the message a deeper meaning. The possibility of multimodal learning can offer a path towards integrating these diverse signals. But there comes a unique ethical challenge when applying AI with teens. If it is used with over-monitoring, opaque profiling, inaccurate classification, or discriminatory training data, it can

become a source of harm: a system designed to become safer can actually be a source of harm. The American Psychological Association (APA, 2025) has highlighted the need to focus on developmental differences, privacy, bias, and safety in interactions between adolescents and AI and digital systems. The goal should not be to watch everything; it should be to detect proportionately so that it can be reviewed in time by humans. This paper presents a conceptual framework of multimodal approach combining both behavioral and natural language analysis to identify digital risks that occur among the adolescents. It is not claiming empirical performance since no new sample of participants is analysed in the current study. Rather, it creates an architecture and an agenda for its evaluation based on a body of existing scholarship. It poses three questions: (1) What are the behavioural and linguistic signals? (2) How to combine these signals with a focus on probabilistic prediction, rather than judgment; and (3) what safeguards are needed for responsible deployment?

II. LITERATURE REVIEW

2.1 Digital Risks in Adolescent Online Environments

The risk faced by adolescents online is complex. There can be multiple types of risks, such as content based, contact based, conduct based and commercial, and multiple types of risks can occur at the same time. Studies regarding the impact of youth digital media have consistently found that the effects of exposure depend on the platform design, developmental stage, and social settings and individual vulnerability of the youth (Odgers & Jensen, 2020). For AI detection, it is crucial to understand the context of the digital behavior, as it can mean different things to different users.

Cyberbullying is an example of the issue of context. While automated systems might detect insults, threats or negative sentiment, bullying can also be repeated, involve a power imbalance, be socially excluded and can involve multi-agency coordinated behaviour. Therefore, determining bullying from a lexical perspective is not a reliable way to determine a harmful interaction. The same caveats apply when trying to pick up on grooming signs, and there is a gradual progression of trust-building, secretiveness, isolation and escalation that needs to be considered as a context to the seemingly benign words.

2.2 Behavioural Analysis

Behavioural analysis is based on patterns, not content. Examples of features that could be potentially included are: Interaction frequency, Temporal regularity, Communication behaviour changes, Network structure, Reciprocity, Account activity, unusual shifts from baseline communication behaviour. The importance of temporal modelling lies in the fact that digital risks can build up over a period. As opposed to independent observations, sequence-based models can describe relationships between events that occur over time. But behavioural data can be very intrusive. Ongoing gathering can develop rich pictures of adolescents' life and relationships. Thus, the responsible approach should strive to provide data minimisation, aggregation and event-level indicators, where feasible. The intent of behavioural analysis should be to offer context and evidence, rather than to 'reconstruct' a young person's whole digital life.

2.3 Natural Language Analysis

Toxic-language detection, hate-speech, harassment and cyberbullying detection have been extensively used for NLP applications. Transformer-based language models have made significant strides over previous bag-of-words models (Devlin et al., 2019; Liu et al., 2019). The use of LLMs also comes with other concerns about explainability, computational cost, bias, and privacy, while also being able to capture longer dependencies and semantic relationships. In the context of adolescents, language signals are particularly complex because words that are slang, sarcastic, humorous, or used by peers will not be identified. The model trained on the adult or non-local language may also be poorly performing on adolescent vernacular. Language detection needs to provide for uncertainty and to not use single phrases
as evidence of harm.

2.4 Multimodal Artificial Intelligence

The aim of multimodal machine learning is to learn from multiple modalities and capitalise on the complementary information they carry. According to Baltrušaitis et al. (2019), the problems of representation, alignment, reasoning and fusion are key issues in multimodal learning. The two streams in digital-risk detection, behaviour and language, are complementary: Language can give semantic evidence, behaviour can give temporal and relational

evidence. There can be an early, intermediate or late stage in the fusion process. Early fusion: Fusion of features before they are predicted, intermediate fusion: Learning of shared representations, and late fusion: Modality-specific predictions fused. Intermediate or "uncertainty-aware" late fusion might be more suitable for adolescents, as it enables the system to keep the explanations at the modality level and lessen reliance on any one type of signal.

2.5 Existing Gap

There are significant advances mentioned in the literature within each isolated task (e.g., toxic-language detection, cyberbullying classification and cyberbullying behaviour modelling). The research gap is to integrate these capabilities into an architecture with safety goals in mind specifically to adolescent digital environments.

Equally important to a useful system is the consideration of uncertainty, developmental context, privacy, and bias and human decision making. This is addressed in the current framework through a multimodal detection approach and explicit layers of governance.

Table 1. Complementary signals used in the proposed framework

Signal stream	Candidate indicators	Contribution	Key caution
Behavioural	Interaction frequency; timing; changes from baseline; network patterns	Provides temporal and relational context	Minimise collection; avoid universal thresholds
Natural language	Harassment; threats; coercion; grooming-like patterns; self-harm-related language	Provides semantic and contextual cues	Account for slang, sarcasm, dialect and uncertainty
Context and time	Sequence of interactions; available contextual information	Helps interpret signals in combination	Use only relevant, ethically permitted context

III. RESEARCH OBJECTIVES AND QUESTIONS

The overall goal is to create a conceptual multimodal AI system that can identify potential risks in the digital world of adolescents by analyzing their behavior and natural language.

Specific objectives include:

1) Defining behavioral and linguistic indicators for digital-risk detection
2) Developing a multimodal fusion architecture that can accommodate a variety of signals
3) Integrating uncertainty estimation and human review into the decision process
4) Defining privacy, bias and explainability protections
5) Defining an empirical validation agenda for future research.

Research Questions

RQ1: Which behavioural and linguistic indicators can support the early identification of potential digital risks among adolescents?

RQ2: How can multimodal fusion integrate behavioural and linguistic signals to improve contextual interpretation and reduce false-positive detections?

RQ3: What ethical and technical safeguards are required to support the responsible deployment of a multimodal AI framework for adolescent digital-risk detection?

IV. PROPOSED MULTIMODAL AI FRAMEWORK

The proposed design has five layered architectures. The first layer is for undertaking data acquisition in a privacy preserving way. The second takes behavioral attributes. The third deals with natural language processing. The fourth one is the uncertainty-aware multimodal fusion. The fifth creates a set of risk categories, and routes high-confidence/high-severity cases to proportionate human review.

4.1 Data Acquisition and Minimisation

Data should be gathered only when there is a valid purpose for safety and legal and ethical permission. Limited text segment, interaction metadata, temporal activity patterns, and voluntary contextual information are candidate inputs that can be considered. Raw data

should be kept to a minimum and if possible, processed on site or be transformed into privacy preserving representations. Sensitive information is data that does not relate to the safety function should not be stored.

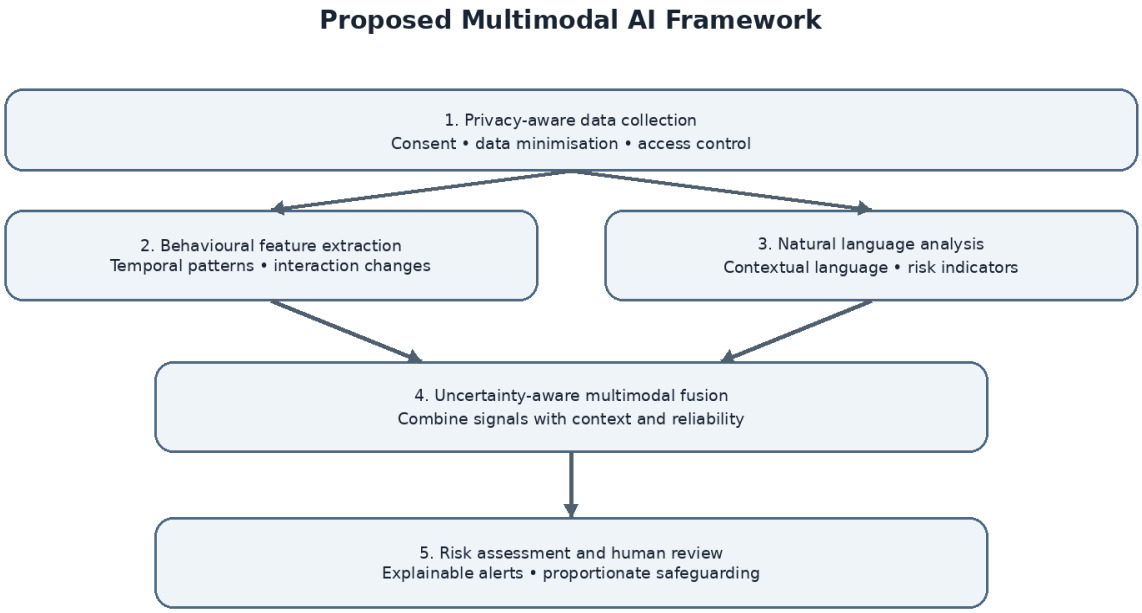


Figure 1. Proposed multimodal AI framework for adolescent digital-risk detection.

4.2 Behavioural Feature Layer

The behavioural layer is for temporal and interaction patterns. Features can include rate of contact, sudden variation from an individual's baseline, multiple interactions with an unknown account, unusual communication timing, rapid increase in communication frequency and network indicators. Normalization should not be done against a universal threshold but rather individual or cohort baseline.

4.3 Natural Language Layer

The language layer identifies indicators related to harassment, threats, coercion, grooming like patterns, sexual exploitation, self-harm related language and other specific risk categories through the use of contextual embeddings and classification models. The model should separate out direct, from those ambiguous, indicators, and offer graduated probabilities instead of binary labels.

4.4 Multimodal Fusion

The fusion layer fuses together behaviour and language representations. A conceptual formulation is R = F (B, L, T, C), B represents behavioural features, L represents language representations, T represents temporal context and C represents available contextual information. The fusion mechanism should not assign equal weight to all inputs, but should take into account the reliability and the uncertainty of the inputs.

4.5 Risk Assessment and Human Review

Should be presented as levels of risk and with confidence intervals and explanations. Low-confidence prediction does not necessarily lead to intervention. Escalation should only be speeded up in high severity patterns but should be in line with pre-defined safeguarding protocols. There should be evidence presented to human reviewers that led to the alert; and there should be a way to reject or override model predictions.

Table 2. Functions and safeguards across the framework layers

Layer	Primary function	Safeguard
1. Data collection	Acquire limited text, metadata and voluntary context	Consent, purpose limitation and retention controls
2. Behavioural features	Represent temporal and interaction patterns	Aggregation and individual/cohort baselines
3. Language analysis	Identify contextual linguistic risk indicators	Uncertainty-aware outputs; no

Layer	Primary function	Safeguard
		single-phrase decisions
4. Multimodal fusion	Combine behaviour, language, time and context	Weight signals by reliability and uncertainty
5. Risk assessment and review	Produce explainable risk levels and route cases for review	Human override and proportionate escalation

V. METHODOLOGY FOR FUTURE EMPIRICAL VALIDATION

The present paper will not be an empirical experiment, but rather a conceptual framework study. Ethically approved datasets with multimodal observations which can be associated with pre-defined risk categories should be used for future validation. Clear inclusion and exclusion criteria, demographic diversity, protocols and procedures for annotation of sensitive material should be included in the construction of the dataset.

5.1 Annotation

Risk categories and contextual severity should be annotated using a multi-annotator process. Specialized training should be provided for annotators and inter-rater agreement should be reported. Any cases that are "ambiguous" should be retained as ambiguous, not classified as a binary class.

5.2 Model Evaluation

Precision, recall, F1 score, area under the receiver operating characteristic curve and calibration measures should be used to evaluate. Precision and false positive rates should be reported together with overall accuracy because false positives can give substantial burdens to the adolescents. Stratification of performance should also be done across the various groups that are pertinent to the research; disparities should be identified.

5.3 Baselines and Ablation

The multimodal system needs to be tested against language-only and behaviour-only baseline. To assess the contribution of individual modalities or temporal components, ablation studies should be performed without the individual modality or temporal component. A separate test set is needed for external validation before generalisability claims can be made.

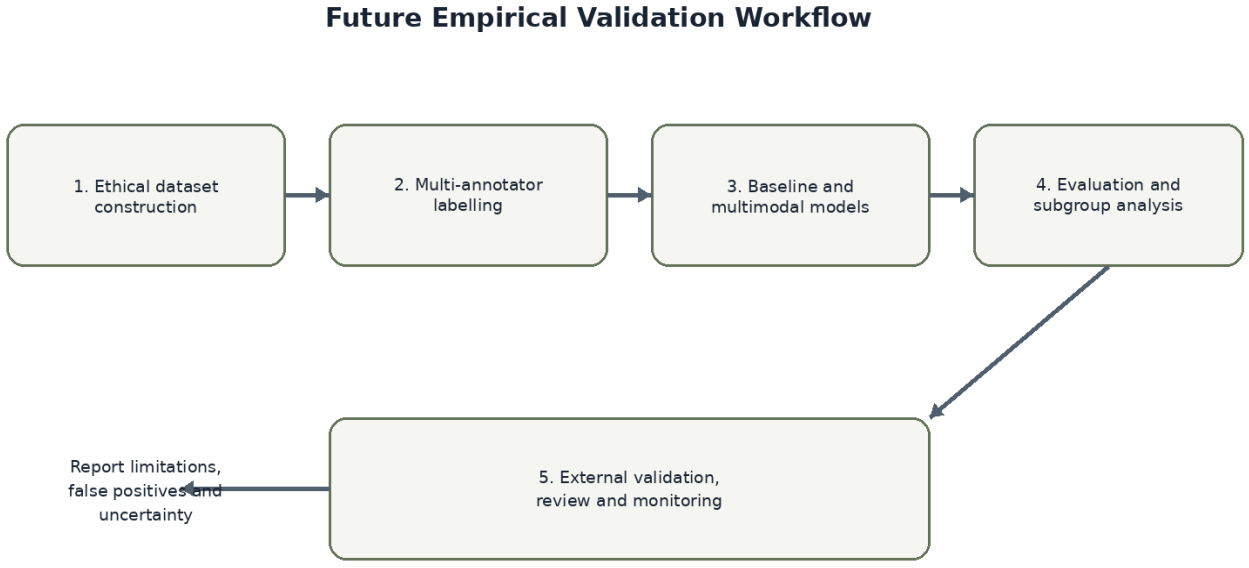


Figure 2. Proposed workflow for future empirical validation.

VI. ETHICAL, PRIVACY AND SAFETY CONSIDERATIONS

The best technical model does not necessarily indicate the safest system. Harms resulting from surveillance or inaccurate classification are exacerbated by developmental vulnerability and power dynamics, especially when the effects of this classification occur in the long-term and are particularly felt by adolescents, and therefore demand extra protection.

Hence, the privacy-by-design approach needs to be incorporated at the architecture level. The collection of data should be minimised, such that it only collects information that is needed for the defined safety goal. Second, retention periods should not be too long, and there should be no ambiguity regarding their definition. Third, access should be granted or denied depending on the role of the user and it should be recorded. Fourth, model outputs should not be considered a proof of misconduct without independent human evaluation. Fifthly, there should be understandable information for adolescents and guardians (as appropriate and legally and ethically permissible). Another major issue that is central is bias. There may be a lack of representation of specific languages, dialects, culture and/or modes of communication in the training data. A model can then over-flag some groups and miss out on risks for others. Subgroup evaluation and continuous monitoring, as well as bias audits, should therefore be included in deployment.

Explainability must go beyond the technical and be practical in nature. The reviewers should be able to see what behavioural and linguistic signals were involved in an alert. There should also be no covert psychological diagnosis in the system. Digital-risk detection is not a diagnosis of depression, suicidality, trauma or any other clinical condition. If serious concerns arise about an individual's wellbeing, the system should alert and pass on to suitably qualified humans, not give a label.

VII. DISCUSSION

The proposed framework addresses the intrinsic complexity of single-modal digital-risk detection, which is that meaning exists across content, behaviour and time. A message is not necessarily bad when it comes alone, but when it comes in a specific relational pattern, it's bad. On the other hand, one could be aggressive words in a peer group that are considered harmless. The evidential basis for human review can, therefore, be enhanced by multimodal contextualization. It also restates detection as a decision support process, and not a decision making one. This separation is important especially for young people. A prediction should not be a predetermined punishment; it should be followed by a proportionate review process. The uncertainty should be represented in the decisions made and the outputs from the models should be considered as probabilistic evidence.

The framework aligns with the overarching principles of responsible AI. Validity, reliability, transparency, explainability, privacy enhancement and fairness are all characteristics that are conducive to trustworthiness and are highlighted by NIST (2023). The OECD (2019) also highlights human-centred values, transparency, robustness, safety and accountability. These principles apply to adolescent digital safety, but how to put them to work to create operational controls can be challenging. Another one is the transformation of the digital environment. The emergence of new platforms, communication methods and generational AI tools has the potential to dramatically change risk circumstances. Static models may thus deteriorate with the passage of time. Lifecycle should include continuous evaluation, drift detection and controlled updating of models. Concurrently, however, 'updating' should not be used as the justification for ongoing adolescent data collection.

VIII. LIMITATIONS AND FUTURE RESEARCH

The main drawback is that there is no empirical data on adolescents to validate the framework yet. As such, it is not possible to claim any predictive accuracy, clinical effectiveness or deployment readiness. The variables proposed are based on the literature and conceptual thinking and need to be tested prospectively.

Further studies are needed to create benchmark datasets that are private, to engage in participatory design with adolescents and with the safeguarding community, and to evaluate different approaches to fusion. It is also important to investigate multilingual and cross-cultural performance, model calibration, adversarial behaviour, distribution shift and impact of false positives on the user. Longitudinal studies would be of even greater benefit as digital risks can be a gradual process.

IX. CONCLUSION

This paper outlines an Artificial Intelligence based approach that aims at the identification of digital risks among adolescent users using a multimodal approach and the complementary analysis of behaviour and natural language. The key is the contextual integration

not mere classification, the central theme of the argument. Language models can give semantic indicators and behavioural signals can give temporal and relational indicators which can be combined to provide a more informative basis for risk assessment. But, don't let multimodal AI be the judge of the safety of teens. The framework thus includes privacy alongside uncertainty estimation, explainability, fairness and human review, alongside of predictive modelling. It is subject to future value – needs to be validated in a prospective way, with data and evidence that it enhances safeguarding while not placing undue surveillance. Multimodal AI can be explored as a supporting tool of the previous, more contextually and more accountably addressing the digital risk.

REFERENCES

[1] American Psychological Association, *Artificial Intelligence and Adolescent Well-Being: An APA Health Advisory*. 2025. [Online]. Available: APA Health Advisory

[2] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2019, doi: 10.1109/TPAMI.2018.2798607.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.

[4] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019. [Online]. Available: arXiv:1907.11692

[5] National Institute of Standards and Technology (NIST), *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, 2023, doi: 10.6028/NIST.AI.100-1.

[6] Organisation for Economic Co-operation and Development (OECD), *Recommendation of the Council on Artificial* Intelligence, 2019.

[7] C. L. Odgers and M. R. Jensen, "Annual research review: Adolescent mental health in the digital age: Facts, fears, and future directions," *Journal of Child Psychology and Psychiatry*, vol. 61, no. 3, pp. 336–348, 2020, doi: 10.1111/jcpp.13190.

[8] S. Livingstone, M. Stoilova, and A. Kelly, "Cyberbullying: Incidence, trends and consequences," in *Handbook of Adolescent Digital Media Use*, J. N. Nesi *et al*., Eds. Routledge, 2019.

[9] S. Zhao, D. Cai, H. Chen, *et al*., "Multimodal machine learning for social media and online safety: A review," *ACM Computing Surveys*, 2021.

[10] L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, B. Schafer, P. Valcke, and E. Vayena, "AI4People—An ethical framework for a good AI society," *Minds and Machines*, vol. 28, pp. 689–707, 2018, doi: 10.1007/s11023-018-9482-5.

[11] UNICEF, *Policy Guidance on AI for Children*, 2021.

[12] UNESCO, *Recommendation on the Ethics of Artificial Intelligence*, 2021.

[13] World Health Organization, *Ethics and Governance of Artificial Intelligence for Health*. WHO, 2021.

[14] European Union Agency for Fundamental Rights, *Getting the Future Right: Artificial Intelligence and Fundamental Rights*. Publications Office of the European Union, 2022.

[15] European Commission, *Ethics Guidelines on Trustworthy AI*, 2024.

[16] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023.

[17] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[19] I. D. Raji, A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, and P. Barnes, "Closing the AI accountability gap," in *Proc. FAT* 2020*, 2020, pp. 33–44, doi: 10.1145/3351095.3372873.

[20] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, T. Gebru, and M. Hardt, "Model cards for model reporting,"

in *Proc. FAT* 2019*, 2019, pp. 220–229, doi: 10.1145/3287560.3287596.

[21] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daumé III, and K. Crawford, "Datasheets for datasets," *Communications of the ACM*, vol. 64, no. 12, pp. 86–92, 2021, doi: 10.1145/3458723.

[22] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, *et al*., "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021. [Online]. Available: arXiv:2108.07258

[23] A. Abid, M. Farooqi, and J. Zou, "Persistent anti-Muslim bias in large language models," in *Proc. AAAI/ACM Conf. AI, Ethics, and Society (AIES)*, 2021, pp. 298–306, doi: 10.1145/3461702.3462624.

[24] A. Caliskan, J. J. Bryson, and A. Narayanan, "Semantics derived automatically from language corpora contain human-like biases," *Science*, vol. 356, no. 6334, pp. 183–186, 2017, doi: 10.1126/science. aal4230.

[25] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in *Proc. ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD)*, 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778.

[26] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[27] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017. [Online]. Available: arXiv:1702.08608

[28] D. Shin, "The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI," *International Journal of Human-Computer Studies*, vol. 146, Art. no. 102551, 2021, doi: 10.1016/j.ijhcs.2020.102551.

[29] S. Amershi, D. Weld, M. Vorvoreanu, A. Fourney, F. Ngan, Q. Baker, P. Collisson, *et al*., "Guidelines for human-AI interaction," in *Proc. CHI Conf. Human Factors in Computing Systems (CHI)*, 2019, pp. 1–13, doi: 10.1145/3290605.3300233.

[30] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of AI ethics guidelines," *Nature Machine Intelligence*, vol. 1, pp. 389–399, 2019, doi: 10.1038/s42256-019-0088-2.

[31] D. Leslie, *Understanding Artificial Intelligence Ethics and Safety*. The Alan Turing Institute, 2019.

[32] European Commission High-Level Expert Group on Artificial Intelligence, *Ethics Guidelines for Trustworthy AI*. European Commission, 2019.

[33] Organisation for Economic Co-operation and Development (OECD), *OECD AI Principles*, 2024.

[34] Council of Europe, *Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law*, 2024.

[35] United Nations, *Governing AI for Humanity: Final Report of the UN Secretary-General's High-Level Advisory Body on Artificial Intelligence*. United Nations, 2024.

[36] UNICEF, *Policy Guidance on AI for Children*, 2020.

[37] S. Livingstone and A. Third, "Children and young people's rights in the digital age," *New Media & Society*, vol. 19, no. 5, pp. 657–670, 2017, doi: 10.1177/1461444816686318.

[38] d. boyd, *It's Complicated: The Social Lives of Networked Teens*. Yale University Press, 2014.

[39] J. W. Patchin and S. Hinduja, "Measuring cyberbullying: Implications for research," *Aggression and Violent Behavior*, vol. 23, pp. 69–74, 2015, doi: 10.1016/j.avb.2015.05.013.

[40] R. M. Kowalski, G. W. Giumetti, A. N. Schroeder, and M. R. Lattanner, "Bullying in the digital age: A critical review and meta-analysis," *Psychological Bulletin*, vol. 140, no. 4, pp. 1073–1137, 2014, doi: 10.1037/a0035618.

[41] I. Zych, R. Ortega-Ruiz, and R. Del Rey, "Systematic review of theoretical studies on bullying and cyberbullying," *Aggression and Violent Behavior*, vol. 23, pp. 1–21, 2015, doi: 10.1016/j.avb.2015.04.001.

[42] A. E. Marwick and d. boyd, "Networked privacy: How teenagers negotiate context in social media," *New Media & Society*, vol. 16, no. 7, pp. 1051–1067, 2014, doi: 10.1177/1461444814543995.

[43] d. boyd and K. Crawford, "Critical questions for big data," *Information, Communication & Society*, vol. 15, no. 5, pp. 662–679, 2012, doi: 10.1080/1369118X.2012.678878.

[44] S. Zuboff, *The Age of Surveillance Capitalism*. PublicAffairs, 2019.

[45] N. Selwyn, *Should Robots Replace Teachers?* Polity, 2019.

[46] B. Williamson and R. Eynon, "Historical threads, missing links, and future directions in AI in education," *Learning, Media and Technology*, vol. 45, no. 3, pp. 223–235, 2020, doi: 10.1080/17439884.2020.1790782.

[47] W. Holmes, M. Bialik, and C. Fadel, *Artificial Intelligence in Education: Promises and Implications for Teaching and Learning*. Center for Curriculum Redesign, 2019.

[48] E. Kasneci, K. Sessler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier, S. Krusche, G. Kutyniok, T. Michaeli, C. Nerdel, F. Pfeiffer, O. Poquet, M. Sailer, A. Schmidt, T. Seidel, M. Stadler, and G. Kasneci, "ChatGPT for good? On opportunities and challenges of large language models for education," *Learning and Individual Differences*, vol. 103, Art. no. 102274, 2023, doi: 10.1016/j.lindif.2023.102274.