Jurnal Unitek
Vol.18 No.1 Januari - Juni 2025
e-ISSN : 2580-2582, p-ISSN: 2089-3957

Etika Dan Bias Dalam LLM: Tanggung Jawab Sosial Atas Kecerdasan Buatan Generatif

Dzulchan Abror [1], Rousyati[2]
[1)] Program Studi Teknologi Komputer, Kampus Kota Tegal, Universitas Bina Sarana Informatika
[2)] Program Studi Sistem Informasi, Kampus Kota Tegal, Universitas Bina Sarana Informatika
Jl. Sipelem No. 22 Kraton, Kota Tegal, Jawa Tengah
Email: e-mail: [1]**dzulchan.dza@bsi.ac.id, [2]rousyati.rou@bsi.ac.id**

ABSTRAK

Large Language Models (LLM) sebagai bagian dari kecerdasan buatan generatif telah membawa lompatan besar dalam pemrosesan bahasa alami. Namun, sejumlah studi menunjukkan bahwa LLM tidak bebas dari bias dan potensi diskriminasi, baik dalam data pelatihan maupun desain modelnya. Penelitian ini bertujuan untuk menganalisis persoalan etika dan tanggung jawab sosial dalam pengembangan dan penerapan LLM, dengan fokus pada isu bias dan dampaknya terhadap keadilan sosial. Metode yang digunakan adalah *narrative literature review* terhadap literatur akademik, laporan kebijakan, dan dokumen industri yang relevan. Hasil kajian menunjukkan bahwa bias dalam LLM bersifat sistemik dan multidimensional, serta belum diimbangi oleh mekanisme tata kelola yang memadai. Penelitian ini merekomendasikan transparansi, audit etis, pelibatan multi *stakeholder*, dan pendekatan riset partisipatif sebagai strategi mitigasi. Dengan demikian, pengembangan LLM ke depan harus berpijak tidak hanya pada efisiensi teknologis, tetapi juga pada prinsip keadilan, akuntabilitas, dan inklusivitas.

Kata kunci: *Large Language Models,* bias, etika, tanggung jawab sosial*, narrative literature review*.

ABSTRACT

Large Language Models (LLMs), as part of generative artificial intelligence, have brought significant advancements in natural language processing. However, numerous studies indicate that LLMs are not free from bias and potential discrimination, whether in training data or model design. This study aims to analyze the ethical and social responsibility issues in the development and deployment of LLMs, focusing on bias and its implications for social justice. The method employed is a narrative literature review of academic literature, policy reports, and relevant industry documents. Findings reveal that bias in LLMs is systemic and multidimensional, and current governance mechanisms are inadequate. This research recommends transparency, ethical audits, multi-stakeholder engagement, and participatory research approaches as mitigation strategies. Hence, future LLM development must prioritize not only technological efficiency but also justice, accountability, and inclusiveness..

Keywords: *Large Language Models, bias, ethics, social responsibility, narrative literature review.*

Pendahuluan

Kecerdasan buatan atau *artificial intelligence* saat ini merupakan salah satu bidang teknologi yang mengalami perkembangan sangat cepat. AI saat ini banyak diaplikasikan dalam segala bidang: kesehatan, pemasaran, pendidikan, contohnya pada media pembelajaran yang berbasis *artificial intelligence* (Sari & Mahmud, 2024). Begitu juga dengan perkembangan salah satu sub cabang pada AI yaitu Large Language Models (LLM) seperti GPT, BERT, dan LLaMA telah mengubah cara manusia berinteraksi dengan teknologi berbasis bahasa. Model ini digunakan dalam berbagai sektor seperti pendidikan, kesehatan, hukum, dan media (Brown et al., 2020). Namun, kemampuan LLM dalam memahami dan menghasilkan teks menyembunyikan tantangan etis yang signifikan, terutama terkait bias dan ketidakadilan representasional. Ketika dilatih pada data yang mencerminkan ketimpangan sosial, LLM berisiko memperkuat stereotip dan diskriminasi (Bender et al., 2021). Fenomena bias ini menimbulkan pertanyaan fundamental mengenai tanggung jawab moral dan sosial dalam desain, pengembangan, dan penerapan teknologi AI. Jika tidak ditangani secara sistemik, bias algoritmik berpotensi memperparah ketimpangan struktural, khususnya di masyarakat rentan secara sosial dan ekonomi (Gordon, 2019).

Penelitian ini bertujuan untuk mengevaluasi bagaimana bias muncul dalam LLM dan bagaimana isu-isu tersebut diposisikan dalam kerangka etika serta tanggung jawab sosial. Dengan pendekatan *narrative literature review*, penelitian ini mengidentifikasi pola bias, mengevaluasi implikasi sosialnya, dan membahas tanggung jawab kolektif pengembang, institusi, serta pembuat kebijakan.

Penelitian ini memberikan kontribusi orisinal melalui tiga aspek utama. Pertama, merumuskan tinjauan lintas disiplin yang mengintegrasikan perspektif etika, sosioteknis, dan kebijakan dalam konteks yang masih jarang dikaji, yaitu negara-negara Global South. Kedua, menyusun klasifikasi kesenjangan riset aktual dan menawarkan arah kebijakan publik berbasis prinsip *Accountable* AI. Ketiga, mengusulkan kerangka evaluasi etika yang dapat diadaptasi secara kontekstual di luar domain Barat. Dengan pendekatan tersebut, artikel ini bertujuan mendorong pengembangan sistem AI yang tidak hanya canggih secara teknis, tetapi juga adil, inklusif, dan bertanggung jawab secara sosial.

Metode Penelitian

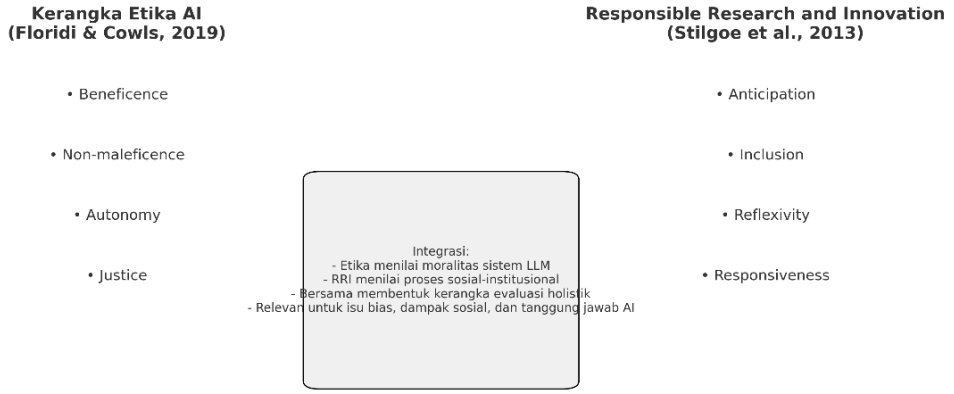
Penelitian ini menggunakan kerangka konseptual yang terdiri atas tiga pendekatan: bias algoritmik, etika teknologi, dan tanggung jawab sosial. Pertama, bias algoritmik dalam LLM sering muncul dari proses pelatihan model dengan data berskala besar yang tidak seimbang. Bias dapat berbentuk *representational bias, allocative bias,* dan *systemic bias* (Mehrabi et al., 2021). *Representational bias* terlihat, misalnya, dalam stereotip gender atau ras yang dihasilkan oleh model (Lucy & Bamman, 2021)(Bender et al., 2021).

Kedua, dari sisi etika teknologi, pengembangan LLM harus tunduk pada prinsip moral seperti keadilan, transparansi, dan non-maleficence atau kewajiban yang tidak menimbulkan bahaya. Berbagai pendekatan etika seperti utilitarianisme, deontologi, dan keadilan distributif dapat digunakan untuk menilai konsekuensi dari LLM (Floridi & Cowls, 2022). Ketiga, tanggung jawab sosial dalam pengembangan teknologi menekankan pentingnya pelibatan publik dan penerapan prinsip *responsible innovation* (Stilgoe et al., 2013). Mekanisme seperti model *cards* (Mitchell et al., 2019)

dan regulasi seperti EU AI Act menjadi instrumen untuk menjamin akuntabilitas sosial teknologi.

Pendekatan *narrative literature review* digunakan untuk mengintegrasikan perspektif etis, teknis, dan sosial terkait bias dalam LLM. Literatur diambil dari jurnal terindeks Scopus dan Web of Science, prosiding ACM/IEEE, serta dokumen kebijakan dari lembaga terpercaya. Total 62 publikasi ditinjau, lalu diseleksi menjadi 20 studi utama berdasarkan relevansi tematik dan dampak akademik. Fokus kajian adalah publikasi dari tahun 2018 hingga 2024 untuk memastikan relevansi dengan perkembangan LLM modern seperti GPT-3, GPT-4, PaLM, dan LLaMA. Namun, literatur klasik yang relevan dengan teori etika dan studi sosioteknis juga disertakan untuk memperkaya konteks historis dan konseptual, dengan strategi pencarian literatur menggunakan kata kunci, seperti "*large language models*", "*bias in AI*", "*ethical AI*", "*generative AI*", "*social responsibility in AI*", dan "*algorithmic fairness*".

Kriteria inklusi mencakup artikel jurnal peer-reviewed dan prosiding konferensi bereputasi dan studi yang secara eksplisit membahas bias, etika, dan tanggung jawab sosial dalam konteks LLM. Artikel opini tanpa dasar kuat dan studi teknis murni tanpa refleksi sosial-etis dikecualikan. Analisis dilakukan secara tematik dengan kerangka dari Floridi & Cowls (Floridi & Cowls, 2022) dan Stilgoe et al. (Stilgoe et al., 2013) untuk menilai keselarasan antara prinsip etika dan implementasi teknologi.



Gambar 1. Kerangka dari Floridi & Cowls dan Stilgoe et al.

Untuk memperkuat validitas pendekatan naratif yang digunakan, Tabel 1 berikut merangkum literatur utama yang dianalisis dalam penelitian ini, termasuk fokus kajian, pendekatan metodologis, dan kontribusinya terhadap tema etika dan bias dalam LLM.

Tabel 1. Ringkasan literatur terkait bias, etika, dan tanggung jawab sosial dalam large language models

No	Penulis &tahun	Fokus kajian	Pendekatan/metode	Kontribusi utama
1	Baumeister & Leary (1997)	Teknik penulisan *narrative review*	Teoretis	Memberikan panduan tentang penyusunan ulasan literatur naratif
2	Bender et al. (2021)	Bahaya model bahasa besar dalam AI	Kajian kritis	Menggugah kesadaran tentang risiko bias dan skalabilitas LLM

No	Penulis &tahun	Fokus kajian	Pendekatan/metode	Kontribusi utama
3	Birhane & Prabhu (2021)	Hegemoni data dan dominasi Global North dalam AI	Kajian sosioteknis	Mengkritisi dampak politik dan ekonomi dari dataset raksasa
4	Blodgett et al. (2020)	Definisi dan praktik evaluasi bias di NLP	Survei kritis	Menekankan perlunya konsistensi dalam definisi dan metodologi evaluasi bias
5	Brown et al. (2020)	Kemampuan *few-shot* learning LLM	Eksperimen GPT-3	Menjadi pijakan penting dalam pengembangan model bahasa generatif
6	Crawford (2021)	Politik dan dampak ekologi dari AI	Kajian etnografis dan reflektif	Menunjukkan sisi tersembunyi dari sistem AI besar, termasuk eksploitasi data
7	Devlin et al. (2019)	Arsitektur BERT dan pre-training untuk NLP	Metode eksperimen	Memberikan dasar arsitektur model yang banyak digunakan sebagai baseline
8	Eubanks (2018)	Ketidakadilan algoritmik dalam kebijakan publik	Kajian kebijakan dan kasus	Menyoroti efek diskriminasi pada kelompok miskin melalui sistem otomatisasi
9	Ferrari (2015)	Struktur dan praktik penulisan naratif literatur	Panduan penulisan akademik	Menekankan pentingnya struktur dan koherensi dalam narrative review
10	Floridi & Cowls (2019)	Kerangka prinsip etika AI	Normatif-teoretis	Merumuskan lima prinsip utama etika AI
11	Gebru et al. (2018)	Dokumentasi dataset melalui *datasheets*	Proposal dokumentasi	Meningkatkan transparansi dan akuntabilitas dalam dataset AI
12	Jasanoff (2004)	Ko-produksi sains dan nilai sosial	Teori STS (Science and Technology Studies)	Menghubungkan pengembangan teknologi dengan struktur sosial
13	Jobin et al. (2019)	Lanskap global pedoman etika AI	Analisis isi dokumen	Mencatat keberagaman dan konsensus dalam etika AI secara internasional
14	Lucy & Bamman (2021)	Representasi gender dalam output LLM	Eksperimen dan analisis linguistik	Membuktikan bahwa LLM mereproduksi stereotip gender
15	Mehrabi et al. (2021)	Taksonomi bias dalam sistem AI	Survei literatur	Klasifikasi jenis bias dan rekomendasi mitigasi
16	Mitchell et al. (2019)	Model Cards untuk pelaporan model AI	Proposal dokumentasi	Memberikan standar evaluasi dan pelaporan sistem AI secara terbuka
17	Mittelstadt (2019)	Kritik terhadap prinsip etika yang tidak implementatif	Analisis normatif	Menyatakan bahwa prinsip etika harus ditunjang mekanisme praktik
18	Noble (2018)	Diskriminasi algoritmik di mesin pencari	Studi kasus dan analisis kritis	Mengungkap bias ras dan gender dalam sistem informasi digital

No	Penulis &tahun	Fokus kajian	Pendekatan/metode	Kontribusi utama
19	Sheng et al. (2019)	Bias dalam teks hasil generasi model bahasa	Eksperimen NLP	Menunjukkan bias eksplisit dalam deskripsi pekerjaan berdasarkan gender
20	Stilgoe et al. (2013)	Framework inovasi yang bertanggung jawab	Kerangka konseptual	Mengusulkan keterlibatan masyarakat dan refleksi nilai dalam inovasi teknologi

Hasil dan Pembahasan

Empat temuan utama dari tinjauan literatur adalah :

1. LLM memperkuat bias sosial

   Studi menunjukkan bahwa LLM seperti GPT-3 dan PaLM secara sistematis memperkuat stereotip berbasis gender, ras, dan agama. (Lucy & Bamman, 2021) menemukan bahwa model cenderung mengasosiasikan profesi tinggi dengan laki-laki dan pekerjaan domestik dengan perempuan. Sebagai produk dari data internet yang bersifat tidak seimbang secara sosiokultural, LLM bukan hanya memantulkan realitas sosial yang bias, tetapi memperkuatnya melalui otomatisasi dan penyebaran ulang secara luas. Ini menciptakan risiko reproduksi ketidakadilan dalam sistem pendidikan, peradilan, dan tenaga kerja (Bender et al., 2021).
2. Transparansi dalam pengembangan LLM masih sangat terbatas.

   Kurangnya dokumentasi terbuka tentang data pelatihan dan arsitektur model menyebabkan hambatan besar terhadap akuntabilitas. Banyak LLM dikembangkan oleh entitas privat dengan pertimbangan komersial, sehingga tidak tunduk pada standar audit publik. Padahal, menurut (Mitchell et al., 2019), dokumentasi seperti model cards dan data sheets for datasets penting untuk menilai potensi bahaya model, termasuk bias sistemik. Tanpa keterbukaan, masyarakat tidak memiliki sarana untuk mengajukan keberatan atau koreksi terhadap hasil yang merugikan.
3. Kesenjangan antara kerangka etika dan praktik implementasi

   Meski banyak organisasi mengadopsi prinsip etika AI (*fairness, accountability, transparency, non-maleficence*), dalam praktiknya prinsip ini sulit dioperasionalkan. (Floridi & Cowls, 2022) menyatakan bahwa etika AI cenderung bersifat deklaratif, bukan transformasional. Banyak pengembang mengutamakan performa teknis seperti fluency atau perplexity, sementara mitigasi bias dan pengujian fairness justru dikesampingkan. Hal ini menunjukkan adanya kesenjangan antara nilai normatif dan insentif praktis dalam ekosistem AI.
4. Perlunya pendekatan partisipatif dan interdisipliner dalam desain dan evaluasi LLM

   Pendekatan teknis semata (seperti *debiasing embeddings* atau *fine-tuning* dengan data seimbang) terbukti tidak cukup. Peneliti seperti (Jasanoff, 2004) dan (Stilgoe et al., 2013) menekankan pentingnya keterlibatan pemangku kepentingan dalam merancang teknologi secara reflektif. Model LLM seharusnya dikembangkan melalui pendekatan partisipatif yang mencakup pengguna akhir, kelompok rentan, regulator, dan ilmuwan sosial. Inilah esensi dari *responsible innovation*, yakni menyeimbangkan antara kemajuan teknologi dan nilai-nilai publik.

Hasil kajian menunjukkan bahwa bias dalam LLM tidak dapat dilihat hanya sebagai persoalan teknis, melainkan juga politis dan sosial. Kecenderungan pasar dan dominasi perusahaan teknologi besar memperumit upaya reformasi etis. Selain itu, literatur juga menyoroti keterbatasan alat evaluasi bias saat ini, yang sering tidak mencerminkan keragaman konteks budaya dan bahasa di luar dunia Barat (Blodgett et al., 2020).

Kebutuhan akan pendekatan interdisipliner menjadi semakin mendesak. Perspektif etika harus dipadukan dengan hukum, antropologi, studi kebijakan, dan hak asasi manusia. Hal ini penting agar mitigasi bias tidak menjadi aktivitas simbolik, tetapi benar-benar menciptakan sistem AI yang adil dan bertanggung jawab secara sosial.

Penelitian ini mengusulkan kerangka evaluasi etika yang dapat diadaptasi dalam konteks non-Barat. Kerangka ini mempertimbangkan keunikan sosial, budaya, dan linguistik yang sering diabaikan dalam evaluasi AI arus utama.

Tabel 2. Kerangka evaluasi etika kontekstual

Dimensi evaluasi	**Konteks non-barat**	**Indikator evaluasi**	**Metode pendekatan**
Representasi sosial	Minoritas lokal, kelompok adat	Kehadiran kelompok dalam data pelatihan & output model	Analisis corpus + FGD lokal
Keadilan linguistik	Bahasa daerah & idiom lokal	Penggunaan idiom & variasi bahasa yang inklusif	Evaluasi linguistik tematik
Partisipasi publik	Keterlibatan komunitas & masyarakat	Keterlibatan dalam desain, pelatihan, dan pengawasan model	Audit partisipatif & survei
Transparansi	Akses informasi di luar teknokratik	Ketersediaan dokumentasi dalam bahasa lokal & akses terbuka	Pelaporan publik & lokalisasi

Kerangka ini bersifat fleksibel dan dapat dikembangkan lebih lanjut dalam studi empiris. Tujuannya adalah untuk memastikan bahwa evaluasi etika terhadap LLM tidak bersifat universalistik dan bias terhadap nilai-nilai Barat, melainkan mempertimbangkan realitas sosial di berbagai konteks global.

Simpulan

Penelitian ini menunjukkan bahwa Large Language Models (LLM), sebagai bentuk kecerdasan buatan generatif, tidak terlepas dari bias sosial, kultural, dan politis yang tertanam dalam data dan struktur pengembangannya. Dengan menggunakan pendekatan *narrative literature review*, kajian ini mengidentifikasi bahwa LLM secara aktif mereproduksi ketimpangan melalui representasi yang bias terhadap gender, ras, agama, dan kelompok rentan lainnya. Ketiadaan transparansi dalam pengembangan model, ditambah lemahnya implementasi prinsip etika, memperbesar potensi kerugian sosial yang ditimbulkan oleh teknologi ini.

Selain itu, terdapat kesenjangan yang signifikan antara wacana etika yang berkembang di level normatif dan praktik nyata dalam pengembangan serta penerapan LLM di industri. Meskipun berbagai inisiatif etika AI telah diluncurkan, banyak yang bersifat simbolis dan kurang mampu menghadirkan perubahan struktural yang diperlukan. Oleh karena itu, bias dalam LLM bukan hanya persoalan teknis, tetapi juga masalah tanggung jawab sosial dan keadilan epistemik yang memerlukan pendekatan interdisipliner dan partisipatif.

Daftar Pustaka

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Blodgett, S. L., Barocas, S., Daumé, H., & Wallach, H. (2020). Language (Technology) is power: A critical survey of ⇜bias" in NLP. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, *c*, 5454–5476. https://doi.org/10.18653/v1/2020.acl-main.485

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *2020-Decem*.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, *1*(Mlm), 4171–4186.

Floridi, L., & Cowls, J. (2022). A unified framework of five principles for AI in society. *Machine Learning and the City: Applications in Architecture and Urban Design*, *July*, 535–545. https://doi.org/10.1002/9781119815075.ch45

Gordon, F. (2019). Virginia Eubanks (2018) Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor. New York: Picador, St Martin's Press. *Law, Technology and Humans*, *1*(1), 162–164. https://doi.org/10.5204/lthj.v1i0.1386

Jasanoff, S. (2004). States of knowledge: The co-production of science and the social order. In *States of Knowledge: The Co-Production of Science and the Social Order*. https://doi.org/10.4324/9780203413845

Lucy, L., & Bamman, D. (2021). *Gender and Representation Bias in GPT-3 Generated Stories*. 48–55. https://doi.org/10.18653/v1/2021.nuse-1.5

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, *54*(6). https://doi.org/10.1145/3457607

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, *Figure 2*, 220–229. https://doi.org/10.1145/3287560.3287596

Sari, F., & Mahmud, S. F. (2024). Implementasi Metode Preference Selection Index Dalam Memilih Media Pembelajaran Matematika Barbasis Artificial Intelligence. *Jurnal Unitek*, *17*(1), 141–151. https://doi.org/10.52072/unitek.v17i1.869

Stilgoe, J., Owen, R., & Macnaghten, P. (2013). Developing a framework for responsible innovation. *Research Policy*, *42*(9), 1568–1580. https://doi.org/10.1016/j.respol.2013.05.008