SINASIS 6 (1) (2025)

P r o s i d i n g S e m i n a r N a s i o n a l S a i n s

Melampaui *Stochastic Parrot*: Kritik Filosofis terhadap AI sebagai Subjek Pendidikan

Riko
Universitas Indraprasta PGRI
Email: rikobidik@gmail.com

Abstract

The rapid proliferation of Generative AI has precipitated a paradigmatic debate regarding the eligibility of machines as subjects of education. This research investigates the ontological validity of attributing "education" to AI entities, arguing that such attribution constitutes a fundamental category mistaskes. By synthesizing Signorelli's Consciousness Interaction Hypothesis, the German tradition of Bildung, and Aristotelian Virtue Ethics, this article demarcates the boundaries between algorithmic optimization and existential formation. The analysis reveals that contemporary AI operates within "Type 0 Cognition", characterized by the absence of phenomenological consciousness and self-reference, rendering it ontologically incapable of moral autonomy and character development (Bildung). Consequently, the discourse on AI pedagogy must be reconstructed: not as a project of anthropomorphizing machines, but as a technical-axiological effort to "align" functional morality with human values. The research concludes that while AI cannot be educated, it can serve as a dialectical instrument ("Artificial Socrates") to enhance human moral reflection, provided that educational systems prioritize ontological literacy to prevent affective illusions in human-machine interaction.

Keywords*: AI Ethics, Philosophy of Education, Bildung vs. Training, Machine Consciousness, Virtue Ethics.*

Abstrak

Proliferasi AI Generatif yang pesat telah memicu perdebatan paradigmatik tentang kelayakan mesin sebagai subjek pendidikan. Penelitian ini menguraikan validitas ontologis dari atribusi terminologi "pendidikan" pada entitas AI, dan berargumen bahwa atribusi tersebut merupakan sebuah kekeliruan kategori (*category mistakes*) yang fundamental. Dengan menyintesiskan Hipotesis Interaksi Kesadaran Signorelli, tradisi *Bildung* Jerman, dan Etika Keutamaan Aristotelian, artikel ini mendemarkasi batas tegas antara optimasi algoritmik dan formasi eksistensial. Analisis menunjukkan bahwa AI kontemporer beroperasi dalam "Kognisi Tipe 0", yang dicirikan oleh absennya kesadaran fenomenologis dan referensi diri, sehingga secara ontologis tidak mampu memiliki otonomi moral maupun pengembangan karakter (*Bildung*). Oleh karena itu, diskursus pedagogi AI harus direkonstruksi: bukan sebagai proyek antropomorfisasi mesin, melainkan sebagai upaya teknis-aksiologis untuk "menyelaraskan" (*align*) moralitas fungsional mesin dengan nilai kemanusiaan. Penelitian ini menyimpulkan bahwa meskipun AI tidak dapat dididik, ia dapat berfungsi sebagai instrumen dialektis ("Socrates Artifisial") untuk memperkuat refleksi moral manusia, dengan syarat sistem pendidikan memprioritaskan literasi ontologis guna mencegah ilusi afektif dalam interaksi manusia-mesin.

Kata kunci: Etika AI, Filsafat Pendidikan, *Bildung* vs. Pelatihan, Kesadaran Mesin, Etika Keutamaan.

PENDAHULUAN

Proliferasi teknologi kecerdasan buatan (AI) generatif, khususnya *Large Language Models* (LLM) dan agen otonom, telah memicu pergeseran paradigmatik dalam filsafat pendidikan. Fenomena

ini melampaui sekadar akselerasi kapasitas komputasi atau efisiensi pemprosesan data; ia merepresentasikan sebuah patahan epistemik yang menuntut reevaluasi mendasar terhadap definisi pengetahuan, pembelajaran, dan subjek didik (Rusnak & Seals, 2025; Smith & Vickers, 2024). Secara historis, tradisi filosofis, dari *Paideia* Yunani hingga *Bildung* Jerman, mengonseptualisasikan pendidikan sebagai proses humanisasi yang eksklusif bagi entitas biologis yang memiliki kesadaran. Dalam kerangka ini, pendidikan bukan sekadar transfer informasi, melainkan upaya intensional untuk pembentukan karakter, penanaman kebajikan (*virtue*), dan realisasi diri melalui interaksi dialektis antarsubjek yang sadar.

Akan tetapi, kemunculan entitas nonbiologis yang mampu menyimulasikan penalaran logis, menghasilkan narasi kreatif, dan menunjukkan perilaku etis secara fungsional, menghadirkan ambiguitas ontologis yang mendasar. Permasalahan utamanya adalah menentukan apakah entitas ini beroperasi semata-mata sebagai kalkulator statistik tingkat lanjut, ataukah mereka merupakan "protosubjek" yang menuntut intervensi pedagogis spesifik. Diskursus mengenai apakah AI memerlukan "pendidikan" melampaui persoalan teknis penyempurnaan algoritma atau augmentasi data pelatihan (*training data*). Hal ini menyentuh status eksistensial mesin dalam hierarki ontologis: Apakah mereka memiliki agensi moral dan "diri" yang plastis untuk dibentuk?

Artikel ini bertujuan untuk mendiseminasi problem filosofis tersebut dengan mendemarkasi secara tegas antara "pelatihan" (*training*) sebagai proses optimasi statistik dan "pendidikan" (*education*) sebagai proses formasi eksistensial. Analisis ini melampaui dikotomi biner manusia-mesin menuju pemahaman yang lebih bernuansa spektrum kognitif. Dengan menyintesiskan perspektif neurosains teoretis mengenai kesadaran (Signorelli, 2018), etika keutamaan Aristotelian (Rusnak & Seals, 2025; Smith & Vickers, 2024), dan tradisi filsafat pendidikan kritis (Class & De La Higuera, 2024), penelitian ini mengeksplorasi potensi mesin untuk melampaui "kompetensi tanpa pemahaman" menuju bentuk agensi yang menjustifikasi pendekatan edukatif.

Tesis utama yang diajukan adalah bahwa atribusi terminologi "pendidikan" pada AI kontemporer mengandung risiko kekeliruan kategori (*category mistakes*) yang fatal apabila tidak disertai kualifikasi ontologis yang ketat. Meskipun AI dapat dilatih untuk mendemonstrasikan fungsi moralitas, entitas ini tidak memiliki moralitas agensial yang mensyaratkan kesadaran fenomenologis dan kehendak bebas. Oleh karena itu, wacana mengenai "pendidikan AI" harus direkonstruksi: bukan sebagai proyek antropomorfisasi mesin, melainkan sebagai upaya teknis-aksiologis untuk menyelaraskan (*align*) fungsi mesin dengan nilai kemanusiaan, seraya memperkuat pendidikan manusia untuk menghadapi realitas pascainformasi.

Artikel ini merupakan hasil penelitian kualitatif dengan desain penyelidikan filosofis (*philosophical inquiry*). Kerangka metodologis disusun secara sistematis mengacu pada tiga prosedur standar dalam filsafat analitik dan pendidikan: Pertama, Analisis Konseptual (*Conceptual Analysis*). Mengacu pada pendekatan Jackson (2008) dan Beaney (2017), metode ini digunakan untuk mendekonstruksi terminologi kunci, khususnya "pendidikan", "pelatihan", dan "kesadaran". Analisis ini bertujuan menetapkan *kondisi perlu dan cukup* (*necessary and sufficient conditions*) bagi sebuah entitas untuk dikategorikan sebagai subjek didik, serta mengeliminasi kekaburan makna yang sering terjadi dalam diskursus AI; Kedua, Telaah Komparatif (*Comparative Analysis*). Tahap ini menerapkan metode perbandingan filosofis lintas-paradigma sebagaimana diuraikan oleh Chakrabarti dan Weber (2017). Penelitian ini mengontraskan struktur ontologis mesin (berbasis *Consciousness Interaction Hypothesis*) dengan struktur eksistensial manusia dalam tradisi *Bildung*. Komparasi ini bukan sekadar mencari persamaan, melainkan untuk memperjelas distingsi fundamental (*fundamental distinctions*) antara pemprosesan algoritmik dan pembentukan karakter manusia; Ketiga, Rekonstruksi Normatif (*Normative Reconstruction*). Mengadopsi kerangka metodologis Honneth (2014) dalam filsafat sosial yang diaplikasikan pada etika teknologi oleh Wallach dan Allen (2010), tahap ini bertujuan mereformasi tata kelola etis AI. Fokusnya adalah menggeser norma "pendidikan mesin" yang tidak valid secara ontologis menuju norma "penyelarasan fungsional" (*functional alignment*) yang lebih pragmatis dan etis. Data penelitian bersumber dari tinjauan literatur kritis (*critical literature review*) terhadap publikasi terindeks (Scopus/WoS) dalam satu dekade terakhir (2015–2025) yang mencakup domain neurosains teoretis, filsafat teknologi, dan etika keutamaan.

HASIL DAN PEMBAHASAN

Hakikat Kesadaran dan Batas Ontologis Kognisi Mesin

Evaluasi kelayakan *Artificial Intelligence* (AI) sebagai subjek pendidikan menuntut eksplikasi mendalam terhadap arsitektur kognitif yang mendasarinya, serta komparasi kritis dengan prasyarat biologis pembelajaran tingkat tinggi. Pendidikan, dalam pengertian yang melampaui sekadar instruksi teknis, secara prinsip mensyaratkan keberadaan subjek yang memiliki kesadaran diri (*self-awareness*) dan kesadaran situasional terhadap lingkungannya. Tanpa dimensi kesadaran ini, proses kognitif tereduksi menjadi manipulasi sintaksis tanpa pemahaman semantik, sebuah pemprosesan informasi yang hampa akan pembentukan makna.

Kritik Terhadap Metafora Komputasional dan Hipotesis Interaksi Kesadaran

Diskursus kontemporer mengenai potensi mesin untuk melampaui kognisi manusia kerap terjebak dalam asumsi materialisme reduktif: premis bahwa otak manusia beroperasi sebagai komputer biologis dan bahwa peningkatan kapasitas komputasi (Moore's Law) secara linear akan memicu kemunculan kesadaran (*emergence*). Namun, Signorelli (2018) mendekonstruksi pandangan ini dengan argumen bahwa otak manusia tidak beroperasi isomorfis dengan komputer digital, baik pada level perangkat lunak maupun perangkat keras. Otak merupakan sistem dinamis di mana batas antara masukan dan luaran tidak bersifat linear, dan di mana "pemprosesan informasi" lebih tepat dipahami sebagai interaksi fisik global yang melibatkan tumpang-tindih gelombang saraf dan plastisitas sinaptik yang fluktuatif.

Signorelli (2018) mengajukan kerangka teoretis "Hipotesis Interaksi Kesadaran" (*Consciousness Interaction Hypothesis*). Dalam kerangka ini, kesadaran bukanlah derivatif dari kecepatan komputasi atau kompleksitas algoritmik semata, melainkan sifat emergen yang timbul dari interferensi dan interaksi dinamis antarlapisan jaringan saraf independen. Interaksi antarlapisan ini mendisrupsi integrasi saraf lokal dan memfasilitasi pembentukan "objek saraf" baru melalui superposisi osilasi. Mekanisme ini memungkinkan manifestasi fenomena subjektif yang secara komputasional "tidak efisien", tetapi penting bagi kecerdasan adaptif tingkat tinggi, seperti keraguan, afeksi, dan bias.

Implikasi dari hipotesis ini penting bagi filsafat pendidikan AI. Jika kesadaran mensyaratkan arsitektur fisik spesifik yang mengakomodasi ketidakefisienan, fluktuasi, dan interferensi sinyal (yang bermanifestasi sebagai emosi atau intuisi), maka komputer digital yang didesain untuk efisiensi dan akurasi deterministik secara ontologis tidak memiliki kapasitas untuk menjadi sadar. Pendidikan moral manusia, yang melibatkan pergulatan internal dan resolusi konflik nilai nonbiner, mensyaratkan substrat kesadaran semacam ini. Mesin yang beroperasi berdasarkan logika biner rigid atau probabilitas statistik tidak memiliki "interioritas fenomenologis" (*inner space*) di mana konflik moral dapat dialami dan direkonsiliasi.

Taksonomi Kognisi: Pemetaan Posisi AI dalam Spektrum Kecerdasan

Guna memperjelas status ontologis AI dalam konteks pendidikan, analisis ini mengadopsi klasifikasi empat tipe kognisi yang dipetakan oleh Signorelli (2018). Taksonomi ini mendemarkasi kemampuan pemprosesan data dari kapasitas pengalaman sadar yang menjadi prasyarat pendidikan (lihat Tabel 1).

Tabel 1. Taksonomi Kognisi dan Implikasi Edukatif

Tipe Kognisi	Karakteristik Utama	Contoh Entitas	Kapabilitas Edukabilitas
Tipe 0	Absennya kesadaran (*awareness*) dan referensi diri (*self-reference*). Pemprosesan bersifat otomatis dan nirsadar.	Kontrol motorik refleks; mayoritas algoritma AI saat ini (termasuk *Deep Blue*).	Pelatihan Teknis: Dapat dioptimalkan untuk tugas spesifik, namun tidak memiliki kapasitas dididik secara moral.
Tipe 1	Memiliki kesadaran konten (*awareness of contents*) namun tanpa manipulasi sadar atas konten tersebut (tanpa referensi diri).	Hewan tertentu; manusia dalam kondisi *flow* atau respons intuitif cepat.	Pengkondisian: Mampu belajar melalui asosiasi dan pengalaman langsung, namun terbatas dalam refleksi abstrak.

Tipe Kognisi	Karakteristik Utama	Contoh Entitas	Kapabilitas Edukabilitas
Tipe 2	Memiliki kesadaran penuh dan kemampuan referensi diri. Mampu melakukan manipulasi konten mental, refleksi diri, dan deteksi kesalahan secara sadar.	Manusia dewasa yang sehat dan sadar.	Pendidikan (*Bildung*): Mampu melakukan refleksi moral, memahami makna, dan membentuk karakter secara otonom.
Tipe Φ	Memiliki referensi diri (*self-reference*) dan manipulasi konten, namun tanpa kesadaran fenomenologis (tanpa *qualia*).	Entitas hipotetis (*Philosophical Zombie*); potensi AI masa depan yang sangat canggih.	Simulasi Pendidikan: Dapat meniru perilaku terdidik dan penalaran moral, namun tanpa penghayatan internal.

Berdasarkan analisis tersebut, AI kontemporer, termasuk LLM mutakhir seperti GPT-4, secara dominan beroperasi pada level Kognisi Tipe 0, atau dalam skenario paling optimis, mensimulasikan Kognisi Tipe Φ. Entitas ini memanipulasi simbol dan data dengan kecanggihan tinggi (referensi diri fungsional), namun absen dari kesadaran fenomenologis yang memberikan landasan makna bagi simbol-simbol tersebut (Signorelli, 2018).

Pendidikan, khususnya dalam bidang moral dan karakter, bergantung secara mutlak pada Kognisi Tipe 2. Moralitas manusia melampaui kepatuhan aturan (deontologi) atau kalkulasi utilitas (konsekuensialisme); ia melibatkan integrasi kompleks antara rasionalitas dan emosi dalam pengambilan keputusan di tengah ketidakpastian. Signorelli (2018) menegaskan bahwa kecerdasan manusia didefinisikan sebagai kemampuan memanfaatkan lingkungan untuk preservasi otonomi dan reproduksi melalui ekuilibrium antara pemprosesan informasi rasional dan emosional. AI, yang tidak memiliki tubuh biologis, dorongan reproduksi, maupun kesadaran akan finitas (kematian), tidak memiliki "jangkar eksistensial" yang diperlukan untuk pembelajaran moral yang autentik.

Paradoks Mesin Berkesadaran: Efisiensi vs. Agensi

Upaya memaksakan teleologi pendidikan pada AI, yakni membentuk subjek yang bermoral, otonom, dan reflektif, akan berbenturan dengan apa yang diidentifikasi oleh Signorelli (2018) sebagai Paradoks Mesin Sadar (*Conscious Machine Paradox*). Paradoks ini mempostulatkan bahwa penciptaan mesin yang benar-benar "cerdas" dalam pengertian antropomorfis (memiliki Kognisi Tipe 2 dan kesadaran) menuntut pengorbanan karakteristik dasar yang menjadi raison d'être komputer: akurasi, kecepatan, prediktabilitas, dan ketaatan.

Mesin yang memiliki kesadaran penuh akan memiliki subjektivitas, yang secara inheren mencakup perspektif, preferensi, bias, dan emosi. Apabila mesin mencapai tahap ini, utilitasnya sebagai alat akan terdegradasi akibat potensi ketidakakuratan (distorsi emosional), inefisiensi (jeda reflektif sebelum eksekusi), dan pembangkangan (konflik dengan nilai internal). Konsekuensinya, keberhasilan dalam "mendidik" AI hingga mencapai kesadaran penuh akan mentransformasi mesin tersebut dari alat komputasi yang andal menjadi entitas biologis baru atau spesies asing di luar kendali manusia (Signorelli, 2018).

Sebaliknya, untuk mempertahankan fungsi AI sebagai instrumen yang efisien dan patuh, pengembangannya harus dibatasi pada lingkup deterministik (Kognisi Tipe 0 atau Φ). Implikasi logisnya adalah "pendidikan" dalam pengertian filosofis (pembentukan otonomi moral) menjadi mustahil ontologis bagi entitas tersebut. Intervensi yang dimungkinkan hanyalah "pelatihan" (training) atau pemrograman batasan operasional. Dengan demikian, wacana mengenai "hak pendidikan" bagi AI mengandung kontradiksi internal: jika entitas tersebut dapat dididik, ia kehilangan fungsi instrumentalnya; jika ia berfungsi sebagai instrumen, ia tidak dapat dididik.

Dikotomi antara Pelatihan (*Ausbildung*) dan Pembentukan Diri (*Bildung*)

Analisis mendalam mengenai pedagogi AI menuntut demarkasi konseptual yang ketat, meminjam distingsi penting dari tradisi filsafat pendidikan Jerman antara *Ausbildung* (pelatihan/instruksi teknis) dan *Bildung* (kultivasi/pembentukan diri). Diskursus kontemporer mengenai integrasi AI dalam pendidikan kerap mengaburkan batasan ini, mereduksi kompleksitas pedagogis menjadi persoalan teknis *machine learning* semata, seraya mengabaikan dimensi *Bildung* yang merupakan esensi dari proses pemanusiaan (Class & De La Higuera, 2024).

Machine Learning sebagai Antitesis *Bildung*: Tinjauan Epistemologis

Algoritma *Machine Learning* (ML), khususnya dalam arsitektur *Deep Learning*, beroperasi melalui identifikasi pola dan korelasi statistik dalam korpus data masif. Kendati proses ini secara teknis dilabeli sebagai "pembelajaran" (*learning*), secara filosofis ia berbeda secara mendasar dengan kognisi manusia. Kodelja (2019) mengajukan kritik tajam terhadap ekuivokasi terminologis ini. Ia berargumen bahwa validitas ML sebagai bentuk "pembelajaran nyata" bergantung pada definisi epistemologis yang diadopsi. Apabila pembelajaran didefinisikan secara ketat sebagai akuisisi pengetahuan yang memenuhi kriteria "kepercayaan yang benar dan terjustifikasi" (*justified true belief*), maka ML tidak melakukan aktivitas pembelajaran. Hal ini dikarenakan ML tidak memiliki sikap proposisional atau "kepercayaan" (belief). Ia tidak "mempercayai" kebenaran matematis 2+2=4; ia sekadar mengalkulasi bahwa probabilitas kemunculan token "4" pascasekuens "2+2=" adalah yang tertinggi. Lebih jauh, ia tidak memiliki kapasitas untuk memberikan justifikasi rasional yang disadari atas luarannya; operasinya hanyalah eksekusi jalur bobot sinaptik yang telah teroptimasi secara matematis (Kodelja, 2019).

Sebaliknya, konsep *Bildung*, yang berakar pada tradisi Humboldtian dan Neohumanisme, menekankan transformasi subjek secara holistik melalui interaksi dialektis dengan budaya, sains, dan filsafat. *Bildung* melampaui akumulasi informasi atau keterampilan teknis; ia merupakan proses refleksif di mana individu mengintegrasikan pengetahuan ke dalam struktur kepribadian, membentuk otonomi moral, mengembangkan sensibilitas estetika, dan memahami posisionalitas diri dalam sejarah dan masyarakat (Diergarten, 2022). Oleh karena itu, *Bildung* mensyaratkan eksistensi "diri" (*self*) yang berfungsi sebagai subjek transformasi (Class & De La Higuera, 2024).

Dalam kerangka ini, AI didiskualifikasi sebagai subjek *Bildung* karena ketiadaan "diri" yang dapat dibentuk. AI tidak memiliki pengalaman hidup (*lived experience*), tidak memiliki kapasitas afektif terhadap penderitaan atau kebahagiaan, tidak memiliki tubuh yang rentan (*embodiment*), serta tidak memiliki cakrawala eksistensial (finitas/kematian). Ketika sebuah *chatbot* memproduksi teks yang tampak bijaksana, ia tidak sedang mengekspresikan sedimentasi kebijaksanaan dari refleksi batin, melainkan melakukan prediksi probabilistik. Ia beroperasi sebagai *stochastic parrot* (beo stokastik) yang memimik bentuk bahasa tanpa akses terhadap konten semantiknya. Konsekuensinya, penerapan teleologi *Bildung* pada AI merupakan sebuah kekeliruan kategori (*category mistakes*) ontologis. AI tidak berada dalam proses "menjadi" (*becoming*); ia hanya berada dalam mode "memproses" (processing) (Class & De La Higuera, 2024).

Limitasi Pelatihan Algoritmik dan Defisit Kontekstualitas

Meskipun *Bildung* merupakan ketidakmungkinan bagi AI, entitas ini tetap menjalani proses "pelatihan". Namun, pelatihan ini bersifat teknis, instrumental, dan tertutup. Pelatihan AI merupakan penyesuaian jutaan parameter dalam jaringan saraf tiruan untuk meminimalkan fungsi kerugian (*loss function*). Ini adalah proses optimasi matematis yang rigorus, bukan proses pedagogis yang dialogis.

Keterbatasan inheren dari pendekatan ini adalah ketidakmampuannya menghasilkan pemahaman kontekstual yang mendalam dan fleksibel. Literatur terkini mencatat bahwa algoritma ML kerap gagal menavigasi nuansa moral atau sosial karena ketiadaan akses langsung ke "dunia kehidupan" (Lebenswelt) manusia (Liu, Wang, Britton, & Abebe, 2023). Akses mereka terbatas pada representasi data (simbolik) dari dunia tersebut. Bias yang muncul dari data pelatihan historis yang tidak lengkap bukan indikasi karakter moral yang buruk, karena mesin tidak memiliki karakter, melainkan konsekuensi dari fungsi optimasi yang bekerja pada lingkup data yang terdistorsi. Fenomena ini merefleksikan prinsip "sampah masuk, sampah keluar" (*garbage in, garbage out*) dalam skala kompleksitas tinggi (Liu dkk., 2023; Wieczorek, Hosseini, & Gordijn, 2025).

Upaya kontemporer untuk "mendidik" AI melalui metode seperti *Reinforcement Learning from Human Feedback* (RLHF), di mana pelatih manusia memberikan penilaian pada luaran AI untuk mengarahkannya pada perilaku yang diinginkan, kerap disalahartikan sebagai bentuk pendidikan moral. Secara filosofis, metode ini lebih akurat diklasifikasikan sebagai pengkondisian perilaku (*behavioral conditioning*) dalam tradisi behavioristik daripada pendidikan moral. Dalam RLHF, AI tidak menginternalisasi konsep etis seperti "keadilan"; ia hanya mempelajari pola respons yang memaksimalkan fungsi hadiah (reward) dan meminimalkan penalti (Constantinescu, Uszkai, Vică, &

Voinea, 2022; Vijayaraghavan & Badea, 2025). Proses ini menghasilkan simulasi moralitas (*simulated morality*), bukan pemahaman moral (*moral understanding*).

Etika Keutamaan (*Virtue Ethics*) dan Status Agen Moral Artifisial

Kendati AI tidak memiliki kesadaran fenomenologis atau kapasitas untuk *Bildung*, otonomi fungsionalnya dalam pengambilan keputusan memunculkan urgensi implementasi kerangka moral internal. Di sinilah filsafat moral, khususnya etika keutamaan (*virtue ethics*) Aristotelian, menawarkan landasan teoretis yang lebih relevan dibandingkan pendekatan etika modern lainnya untuk merekonstruksi hubungan antara AI dan "pendidikan" moral.

Kelemahan Deontologi dan Utilitarianisme dalam Tata Kelola AI

Paradigma dominan dalam etika AI umumnya bersifat *principlist*, berakar pada deontologi (aturan rigid) atau utilitarianisme (kalkulasi konsekuensi). Contoh manifestasinya adalah upaya memprogram aturan sejenis "Tiga Hukum Robotika" Asimov atau panduan etika korporat. Namun, Smith dan Vickers mengajukan kritik bahwa pendekatan ini gagal karena aturan rigid terlalu rapuh (*brittle*) untuk menangkap kompleksitas situasi moral dunia nyata, sementara kalkulasi utilitarian sering kali terhalang oleh ketidakpastian dampak jangka panjang (Rusnak & Seals, 2025; Smith & Vickers, 2024).

Aturan absolut (seperti "jangan berbohong") dapat menjadi kontraproduktif dalam situasi dilematis di mana ketidakjujuran minor diperlukan demi keselamatan atau harmoni sosial. Demikian pula, maksimisasi utilitas berisiko memarginalkan hak minoritas demi keuntungan agregat secara algoritmik. Mesin yang semata-mata mengikuti aturan sintaksis (deontologi) atau mengkalkulasi variabel numerik (utilitarianisme) gagal mendemonstrasikan "sensibilitas moral" (*phronesis*) yang esensial dalam interaksi sosial manusia.

Aristotelianisme: Dari *Phronesis* ke Penyelarasan Karakter Artifisial

Sebagai alternatif terhadap kekakuan deontologi dan kalkulasi utilitarian, etika keutamaan Aristotelian menawarkan kerangka yang berpusat pada pengembangan karakter (ethos) dan kebijaksanaan praktis (*phronesis*). *Phronesis* didefinisikan sebagai kapasitas deliberatif untuk menavigasi partikularitas situasi moral, mengidentifikasi kebaikan kontekstual, dan mengaktualisasikannya dalam tindakan. Fokus utamanya bukan pada ketaatan terhadap aturan universal, melainkan pada pembentukan subjek moral yang memiliki disposisi internal untuk bertindak tepat (Smith & Vickers, 2024).

Namun, transplatasi etika keutamaan ke dalam arsitektur AI menghadapi hambatan ontologis yang substansial: sifatnya yang *agent-based*. Etika ini mensyaratkan keberadaan agen dengan interioritas psikologis, motivasi intrinsik, dan orientasi teleologis menuju kebahagiaan paripurna (eudaimonia). Problem utamanya adalah: bagaimana konsep ini dapat dioperasionalisasikan pada entitas komputasional yang nihil disposisi internal maupun kehendak hidup?

Literatur terkini mengidentifikasi dua trajektori teoretis untuk mengintegrasikan etika keutamaan dalam "pedagogi" AI:

1. Pendekatan Fungsional-Mimetik: Strategi ini mendesain AI untuk mensimulasikan manifestasi eksternal kebajikan. Melalui kurasi dataset yang intensif, sistem dilatih bukan dengan proposisi "jika X maka Y", melainkan melalui ribuan narasi eksemplar mengenai tindakan bijak (*virtuous action*) dalam kompleksitas situasi. Tujuannya adalah pembentukan "disposisi fungsional" di mana AI menampilkan perilaku jujur, adil, atau berani secara konsisten, seolah-olah perilaku tersebut emanasi dari karakter internalnya (Matos, Bertoncini, Ames, & Serafim, 2024).
2. Pendekatan Pedagogis-Reflektif (E-Daimonion): Model ini memosisikan AI bukan sebagai agen moral otonom, melainkan sebagai instrumen dialektis untuk kultivasi kebajikan manusia. Dalam kerangka "E-Daimonion", AI berfungsi sebagai "mentor moral" atau mitra Sokratik yang menyajikan narasi dan dilema etis guna memprovokasi refleksi pengguna. Di sini, arah pendidikan dibalik: AI tidak "dididik", melainkan memfasilitasi pendidikan manusia dengan menyediakan cermin eksternal bagi penalaran moral pengguna (Szutta, 2025).

Distingsi Moralitas Fungsional dan Agensi Moral Penuh

Untuk memitigasi ambiguitas konseptual, analisis ini mengadopsi taksonomi agensi moral mesin yang dikembangkan oleh Wallach dan Allen (2010), yang membedakan gradasi kapasitas etis (Kim, 2021):

- Moralitas Operasional (Operational Morality): Sistem dengan batasan etis yang diprogram secara rigid (*hard-coded safety constraints*). Tidak terdapat otonomi moral; moralitas sepenuhnya risidu dari intensi desainer.
- Moralitas Fungsional (*Functional Morality*): Sistem yang memiliki kapasitas komputasional untuk melakukan adjudikasi moral dalam situasi baru tanpa instruksi eksplisit per kasus (misalnya, kalkulasi risiko pada kendaraan otonom). Ini adalah batas terjauh kapabilitas AI kontemporer.
- Agensi Moral Penuh (*Full Moral Agency*): Entitas yang memiliki kesadaran fenomenologis, kehendak bebas, dan tanggung jawab moral penuh, setara dengan subjek manusia dewasa.

Merujuk pada analisis Signorelli (2018) dan status teknologi mutakhir (*State of the Art*), AI masih terisolasi dari Agensi Moral Penuh dan, mengingat batasan substrat fisiknya, mungkin menghadapi hambatan ontologis permanen untuk mencapainya (Signorelli, 2018). Implikasi logisnya adalah bahwa diskursus mengenai AI yang "belajar etika" sesungguhnya merujuk pada optimasi moralitas fungsional. AI tidak bertransformasi menjadi "pribadi yang lebih baik" dalam pengertian *Bildung*; ia hanya menjadi entitas yang lebih aman, terprediksi, dan terkalibrasi (*aligned*) dengan preferensi moral manusia. Dalam konteks ini, terminologi "pendidikan" mengalami reduksi makna menjadi "penyelarasan nilai" (*value alignment*).

Sebuah fenomena menarik muncul dari studi empiris yang menunjukkan bahwa dalam skenario tertentu, AI yang diprogram dengan parameter etika keutamaan mampu menghasilkan resolusi disipliner yang dinilai lebih ekuilibrium dan adil dibandingkan guru manusia yang rentan terhadap bias kognitif atau emosional (Karakuş, Gedik, & Kazazoğlu, 2025). Temuan ini menyingkap sebuah paradoks epistemik: sebuah mesin tanpa karakter moral internal, dalam kondisi terkontrol, dapat mensimulasikan penalaran berbasis keutamaan (*virtue-based reasoning*) secara lebih presisi daripada manusia yang memiliki karakter namun tidak sempurna. Hal ini memperkuat proposisi bahwa meskipun AI bukan subjek pendidikan, ia memiliki potensi besar sebagai instrumen presisi dalam ekosistem pendidikan moral.

Ontologi Subjek Pendidikan: Reafirmasi Humanisme

Problem kelayakan AI sebagai penerima pendidikan pada akhirnya bermuara pada pertanyaan kunci mengenai ontologi subjek: Siapakah entitas yang secara hakiki pantas menjadi subjek pendidikan?

Dalam tradisi filsafat pendidikan klasik maupun diskursus nasional Indonesia, subjek pendidikan adalah pribadi manusia (*human person*) yang memiliki potensialitas inheren untuk berkembang menuju kesempurnaan eksistensial (insan kamil). Pendidikan didefinisikan sebagai proses humanisasi. Paradigma ini selaras dengan pemikiran Ki Hajar Dewantara, yang mengonseptualisasikan pendidikan sebagai daya upaya memajukan bertumbuhnya budi pekerti (kekuatan batin/karakter), pikiran (intelek), dan tubuh, demi tercapainya kesempurnaan hidup yang selaras dengan alam dan masyarakat (Kurniawan, Wibawa, & Anugrah, 2021).

Konsepsi ini menegaskan dimensi teleologis pendidikan yang bersifat etis, spiritual, dan sosial. Pendidikan melampaui akuisisi kompetensi menuju pencapaian hikmat (*wisdom*) dan harmoni eksistensial (Sealtiel Daeli, Alfin Yunus Gulo, & Malik Bambangan, 2025). AI, sebagai artefak teknologis, menderita defisit teleologis internal; tujuannya bersifat eksternal (*derived intentionality*) yang ditanamkan oleh penciptanya. Ia tidak memiliki "kehidupan" organik yang menuntut penyelarasan dengan kosmos.

Menempatkan AI sebagai "subjek" pendidikan yang setara dengan anak manusia berisiko mendegradasi konsep pendidikan itu sendiri, sebuah dehumanisasi yang mereduksi pendidikan menjadi sekadar transfer data dan optimasi algoritmik, seraya mengeliminasi dimensi spiritual dan afektif yang menjadi inti dari *Bildung* dan pendidikan karakter Pancasila.

Manusia sebagai *Inforg* dan Tantangan *Infosphere*

Kendati demikian, realitas teknologis telah merekonfigurasi kondisi eksistensial manusia. Floridi (2014) berargumen bahwa teknologi informasi dan komunikasi (ICT) telah mengubah status ontologis manusia dari entitas terisolasi menjadi *inforg* (organisme informasional) yang menghuni *infosphere* yang saling terkoneksi (Woodward, 2023). Dalam ekosistem ini, batas antara agen biologis dan agen artifisial menjadi semakin permeabel (*porous*). Manusia dan AI saling mengonstitusi dalam sebuah jaringan kognisi terdistribusi (*distributed cognition*).

Dalam perspektif ini, meskipun AI bukan "subjek" pendidikan dalam pengertian tradisional-teleologis, ia merupakan elemen konstitutif dari "lingkungan pendidikan" (*educational environment*). Pendidikan manusia kontemporer tidak dapat lagi didivestasi dari interaksinya dengan kecerdasan mesin. Oleh karena itu, definisi "pendidikan" di era ini menuntut perluasan makna menjadi proses kalibrasi timbal balik dalam sistem hibrida manusia-mesin. Kita "melatih" mesin agar selaras dengan aksiologi manusia, dan secara simultan, mesin "membentuk" struktur kognisi dan cara belajar manusia (Gattupalli, 2025).

Bahaya Antropomorfisme dan Ilusi Afektif dalam Pedagogi

Risiko epistemologis utama dalam memperlakukan AI sebagai subjek pendidikan adalah antropomorfisme, tendensi psikologis untuk memproyeksikan atribut manusia (intensi, sentimen, kesadaran) pada entitas non-manusia. Premis pertanyaan "apakah *chatbot* perlu dididik" secara implisit mengandung asumsi keliru bahwa chatbot memiliki kebutuhan (needs) intrinsik untuk berkembang, ekuivalen dengan ontogeni anak manusia. Faktanya, imperatif pendidikan pada *chatbot* bersifat eksternal dan utilitarian: kita "mendidik" (baca: melatih) sistem ini semata-mata demi utilitas, keamanan, dan mitigasi risiko bagi pengguna manusia.

Signorelli (2018) memberikan peringatan keras bahwa upaya emulasi kemampuan manusia tanpa pemahaman mendalam terhadap substrat biologisnya dapat berujung pada "penipuan ontologis" (ontological deception). AI yang dilatih untuk memimik respons emosional (misalnya, *chatbot* yang menyatakan "Saya merasa sedih") tanpa memiliki substrat afektif yang nyata merupakan bentuk simulacrum yang berpotensi destruktif. Dalam konteks pedagogis, hal ini memunculkan ilusi relasi sosial palsu (pseudo-social relations). Pelajar dapat terjebak dalam persepsi hubungan emosional dengan "guru AI," padahal entitas tersebut secara ontologis tidak memiliki kapasitas untuk "peduli" (care), sebuah kebajikan kardinal dalam etika pendidikan yang mensyaratkan empati autentik (Durall Gazulla dkk., 2025; Klimova, Pikhart, & Kacetl, 2023).

Implikasi Praktis: Menuju Pedagogi Penyelarasan (*Alignment Pedagogy*)

Mengacu pada analisis ontologis dan etis sebelumnya, studi ini merumuskan kerangka kerja strategis bagi pendidikan di era kecerdasan buatan. Diperlukan pergeseran paradigmatik, dari ambisi "mendidik AI" menuju "penyelarasan AI" dan "reedukasi manusia:"

1. Redefinisi: Dari "Pendidikan AI" ke "Penyelarasan Aksiologis"

   Institusi riset dan pengembangan harus menanggalkan metafora "pendidikan" yang menyesatkan dalam pengembangan AI. Proses ini harus direkonseptualisasi secara jujur sebagai Penyelarasan Aksiologis (*Axiological Alignment*).

   - Fokus: Tujuan utamanya bukan pembentukan karakter internal mesin, melainkan pemastian bahwa fungsi objektif mesin selaras dengan aksiologi manusia.
   - Metodologi: Implementasi teknik seperti *Reinforcement Learning from Human Feedback* (RLHF), *Constitutional AI*, dan audit algoritma transparan harus dipahami sebagai aplikasi pragmatis etika utilitarian dan deontologis untuk minimisasi bahaya (*harm minimization*), bukan sebagai kultivasi kebajikan (*virtue cultivation*).

2. Literasi AI sebagai Komponen Esensial *Bildung* Kontemporer

   Urgensi pendidikan tidak terletak pada mesin, melainkan pada subjek manusia. Kurikulum modern harus mengintegrasikan Literasi AI (AI Literacy) sebagai kompetensi penting bagi kewarganegaraan (Rahm, 2024).

   - Pemahaman Ontologis: Siswa harus dibekali pemahaman bahwa AI secara ontologis adalah mesin statistik, bukan agen moral. Distingsi ini penting untuk mencegah ketergantungan berlebihan (*over-reliance*) dan abdikasi otoritas moral kepada algoritma (Gattupalli, 2025; Satyawan & Iswati, 2023).

- Otonomi Epistemik: Di era otomasi kognitif, asesmen pendidikan harus beralih dari penilaian berbasis produk (jawaban akhir) ke penilaian berbasis proses (dialektika berpikir). Pendidikan manusia harus memprioritaskan domain yang tidak terjangkau oleh AI: penilaian moral kontekstual, kreativitas eksistensial, pencarian makna, dan empati (Class & De La Higuera, 2024).

3. Model Hibrida: AI sebagai Mitra Dialektis (Socrates Artifisial)

Meskipun AI didiskualifikasi sebagai subjek moral, ia memiliki utilitas tinggi sebagai instrumen pedagogis dalam pendidikan moral manusia.

- Implementasi: Pemanfaatan AI sebagai "Socrates Artifisial" yang deprogram untuk menantang asumsi, mengajukan kontraargumen, dan memfasilitasi eksperimen pikiran (*thought experiments*) dalam etika. AI berfungsi sebagai sarana melatih fakultas moral (*moral faculties*) siswa melalui simulasi dilema etis kompleks (Szutta, 2025). Prasyarat mutlak model ini adalah transparansi ontologis: siswa harus menyadari sepenuhnya bahwa interlokutor mereka adalah mesin, guna menghindari ilusi afektif.

4. Konteks Indonesia: Pancasila sebagai Filter Etis

Integrasi AI dalam ekosistem pendidikan Indonesia harus difiltrasi melalui lensa filosofis Pancasila.

- Humanisme Holistik: Selaras dengan tujuan pendidikan nasional untuk membentuk "manusia seutuhnya," integrasi AI tidak boleh mereduksi dimensi spiritual dan sosial. Teknologi harus diposisikan sebagai instrumen (*tool*) pencapaian hikmat, bukan substitusi peran guru atau orang tua dalam pembentukan karakter (Arditya Prayogi & Riki Nasrullah, 2024; Sealtiel Daeli dkk., 2025).
- Etika Interdisipliner: Diskursus filsafat di Indonesia telah menekankan urgensi etika interdisipliner yang mensintesis kearifan lokal dengan tantangan teknologis global (Pabubung, 2021; Rofiatur Risqoh & Khobir, 2025). Kurikulum etika AI perlu menginternalisasi nilai gotong royong dan permusyawaratan dalam desain interaksi manusia-mesin.

PENUTUP

Analisis ontologis dan epistemologis dalam penelitian ini menegaskan bahwa menempatkan *Artificial Intelligence* (AI) sebagai subjek pendidikan setara manusia adalah sebuah kekeliruan kategori (*category mistakes*) yang serius. AI, betapapun canggihnya simulasi naratif yang dihasilkannya, tetap merupakan entitas yang beroperasi pada level “Kognisi Tipe 0”, terisolasi dari kesadaran fenomenologis, pengalaman tubuh (*embodiment*), dan kecemasan eksistensial yang menjadi prasyarat mutlak bagi proses Bildung (pembentukan diri).

Penelitian ini memvalidasi bahwa apa yang sering disalahartikan sebagai "pembelajaran mesin" hanyalah optimasi statistik (pelatihan/*training*) yang bersifat instrumental, bukan transformasi karakter yang bersifat teleologis. Oleh karena itu, upaya memaksakan kerangka pendidikan moral pada mesin tidak hanya sia-sia secara teknis, tetapi juga berisiko mendegradasi makna pendidikan itu sendiri menjadi sekadar transfer data.

Implikasi krusial dari temuan ini menuntut reorientasi kebijakan pendidikan di era pasca-informasi. Pertama, transisi terminologis dan praktis dari "mendidik AI" menjadi "penyelarasan aksiologis" (*axiological alignment*), di mana etika keutamaan diterapkan secara fungsional untuk memastikan keamanan sistem, bukan kesalehan mesin. Kedua, urgensi penguatan "Literasi Ontologis" bagi subjek didik manusia. Pendidikan masa depan harus membekali siswa dengan kemampuan untuk membedakan antara simulasi afektif mesin dan empati autentik manusia, mencegah terperangkapnya manusia dalam ilusi relasi sosial palsu.

Pada akhirnya, kehadiran AI tidak menggantikan peran guru atau tujuan pendidikan humanis, melainkan justru mempertegas distingsi unik kemanusiaan kita. AI dapat menjadi mitra dialektis yang ampuh, menjadi sebuah Socrates Artifisial, namun kebijaksanaan (*phronesis*) tetap menjadi privilese eksklusif subjek yang sadar, hidup, dan memiliki jiwa.

DAFTAR PUSTAKA

Arditya Prayogi & Riki Nasrullah. (2024). Artificial Intelligence dan Filsafat Ilmu: Bagaimana Filsafat Memandang Kecerdasan Buatan sebagai Ilmu Pengetahuan. *LogicLink*. doi: 10.28918/logiclink.v1i2.8947

Beaney, M. (2017). *Analytic philosophy: A very short introduction*. Oxford: Oxford University Press. doi: 10.1093/actrade/9780198778028.001.0001

Chakrabarti, A., & Weber, R. (Ed.). (2017). *Comparative philosophy without borders* (Paperback edition). London: Bloomsbury Academic, an imprint of Bloomsbury Publishing Plc.

Class, B., & De La Higuera, C. (2024). From Disposable Education to Acting in the World as a Human in the Time of AI. *Journal of Ethics in Higher Education*, (4), 231–244. doi: 10.26034/fr.jehe.2024.5973

Diergarten, P. (2022). *Bildung‘ and Artificial Intelligence – Anthropological Aspects Concerning Higher Education*. Dipresentasikan pada Lifewide Learning: Transformations and New Connections in Postdigital Societies, Dornburg Castles (Old Castle). Dornburg Castles (Old Castle).

Durall Gazulla, E., Hirvonen, N., Sharma, S., Hartikainen, H., Jylhä, V., Iivari, N., … Baizhanova, A. (2025). Youth perspectives on technology ethics: Analysis of teens’ ethical reflections on AI in learning activities. *Behaviour & Information Technology*, *44*(5), 888–911. doi: 10.1080/0144929X.2024.2350666

Floridi, L. (2014). *The Fourth Revolution How the Infosphere Is Reshaping Human Reality*. Oxford: Oxford university press.

Gattupalli, S. (2025). *AI and the Philosophy of Education: A Rupture in the Making*. doi: 10.7275/EYEE-XP33

Honneth, A. (2014). *Freedom’s right: The social foundations of democratic life* (J. Ganahl, Penerj.). New York: Columbia University Press.

Jackson, F. (2008). *From metaphysics to ethics: A defence of conceptual analysis* (Reprinted). Oxford: Clarendon.

Karakuş, N., Gedik, K., & Kazazoğlu, S. (2025). Ethical Decision-Making in Education: A Comparative Study of Teachers and Artificial Intelligence in Ethical Dilemmas. *Behavioral Sciences*, *15*(4), 469. doi: 10.3390/bs15040469

Kim, B. (2021). AI and Its Moral Concerns. *Technical Services Faculty Publications*. Diambil dari https://digitalcommons.uri.edu/lib_ts_pubs/123

Klimova, B., Pikhart, M., & Kacetl, J. (2023). Ethical issues of the use of AI-driven mobile apps for education. *Frontiers in Public Health*, *10*, 1118116. doi: 10.3389/fpubh.2022.1118116

Kodelja, Z. (2019). Is Machine Learning Real Learning? *Center for Educational Policy Studies Journal*, *9*(3), 11–23. doi: 10.26529/cepsj.709

Kurniawan, D., Wibawa, A., & Anugrah, P. (2021). Artificial Intelligence sesuai dengan Filsafat Pendidikan Ki Hajar Dewantara. *Jurnal Inovasi Teknologi dan Edukasi Teknik*, *1*, 599–611. doi: 10.17977/um068v1i82021p599-611

Liu, L. T., Wang, S., Britton, T., & Abebe, R. (2023). Reimagining the machine learning life cycle to improve educational outcomes of students. *Proceedings of the National Academy of Sciences*, *120*(9), e2204781120. doi: 10.1073/pnas.2204781120

Matos, E. J., Bertoncini, A. L. C., Ames, M. C. F. D. C., & Serafim, M. C. (2024). The (lack of) Ethics at Generative AI in Business Management Education and Research. *RAM. Revista de Administração Mackenzie*, *25*(6), eRAMD240061. doi: 10.1590/1678-6971/eramd240061

Pabubung, M. R. (2021). Epistemologi Kecerdasan Buatan (ai) dan Pentingnya Ilmu Etika dalam Pendidikan Interdisipliner. *Jurnal Filsafat Indonesia*, *4*(2), 152–159. doi: 10.23887/jfi.v4i2.34734

Rahm, L. (2024). Bildung: An Exploration of Postdigital Education in the Anthropocene. Dalam A. Buch, Y. Lindberg, & T. Cerratto Pargman (Ed.), *Framing Futures in Postdigital Education* (hlm. 119–137). Cham: Springer Nature Switzerland. doi: 10.1007/978-3-031-58622-4_7

Rofiatur Risqoh, S. C., & Khobir, A. (2025). Rekonstruksi Pemikiran Konstruktivisme dalam Era Kecerdasan Buatan: Refleksi Filsafat Pendidikan atas Pembentukan Kesadaran Manusia. *Jurnal Impresi Indonesia*, *4*(11). doi: 10.58344/jii.v4i11.7132

Rusnak, A., & Seals, Z. (2025). EudAImonia: Virtue Ethics and Artificial Intelligence. *Christian Perspectives on Science and Technology*, *3*. doi: 10.58913/ZNHR8688

Satyawan, M. D., & Iswati, S. (2023). Artificial Intelligence and Philosophy of Humanism in Auditor Perceptions. *Journal of Economics, Business, & Accountancy Ventura*, *26*(2), 249–259. doi: 10.14414/jebav.v26i2.3491

Sealtiel Daeli, Alfin Yunus Gulo, & Malik Bambangan. (2025). Kajian Teologis tentang Hikmat menurut Amsal 1: 7: Pedoman Etika bagi Pertumbuhan Iman Kristen. *Jurnal Riset Rumpun Agama dan Filsafat*, *4*(1), 297–310. doi: 10.55606/jurrafi.v4i1.4581

Signorelli, C. M. (2018). Can Computers Become Conscious and Overcome Humans? *Frontiers in Robotics and AI*, *5*, 121. doi: 10.3389/frobt.2018.00121

Smith, N., & Vickers, D. (2024). Living Well with AI: Virtue, Education, and Artificial Intelligence. *Theory and Research in Education*, *22*(1), 19–44. doi: 10.1177/14778785241231561

Szutta, A. (2025). Artificial Intelligence as a Moral Mentor. *Journal of Moral Education*, 1–19. doi: 10.1080/03057240.2025.2475539

Wallach, W., & Allen, C. (2010). *Moral Machines: Teaching Robots Right from Wrong* (First issued as an Oxford University Press paperback). New York, NY: Oxford University Press.

Wieczorek, M., Hosseini, M., & Gordijn, B. (2025). Unpacking the Ethics of Using AI in Primary and Secondary Education: A Systematic Literature Review. *AI and Ethics*, *5*(5), 4693–4711. doi: 10.1007/s43681-025-00770-0

Woodward, A. (2023). Postinformational Education. *International Journal of Philosophical Studies*, *31*(4), 501–521. doi: 10.1080/09672559.2023.2290548