Ontologija Pogreške: Algoritamska Entropija i Arhontska Eksploatacija u Digitalnom Dobu

Autor: Davor Moravek

Sažetak

Ovaj rad predstavlja teorijski okvir za razumijevanje fundamentalnih poremećaja u digitalnom dobu, koristeći sintezu drevnih filozofskih modela i suvremene teorije umjetne inteligencije. Središnja teza jest da digitalni kaos nije monolitan, već se manifestira kroz dvije različite, ali isprepletene strukture pogreške: **Algoritamsku Entropiju** – nenamjerni, amoralni kaos koji proizlazi iz inherentnih tehničkih ograničenja složenih sustava – i **Arhontsku Eksploataciju** – namjernu, zlonamjernu manipulaciju tih sustava od strane aktera koji iskorištavaju njihove slabosti. Analizom ove dualnosti, rad predlaže normativni okvir „Etiku Aše" kao operativni model za dizajn sustava koji inherentno teže istini i provjerljivosti. Konačno, predstavljen je i detaljan istraživački protokol za empirijsku validaciju predloženog okvira u visoko-rizičnoj domeni.

Ključne riječi: ontologija pogreške, algoritamska entropija, arhontska eksploatacija, etika umjetne inteligencije, nadzorni kapitalizam, kauzalno rezoniranje

Abstract

This paper introduces a theoretical framework for understanding the fundamental disruptions of the digital age, using a synthesis of ancient philosophical models and contemporary artificial intelligence theory. The central thesis is that digital chaos is not monolithic but manifests through two distinct yet intertwined structures of error: **Algorithmic Entropy**—unintentional, amoral chaos arising from the inherent technical limitations of complex systems—and **Archontic Exploitation**—the intentional, malicious manipulation of these systems by actors who exploit their weaknesses. By analyzing this duality, the paper proposes the "Ethics of Asha" as a normative framework and an operational model for designing systems that inherently strive for truth and verifiability. Finally, a detailed research protocol for the empirical validation of the proposed framework in a high-stakes domain is presented.

Keywords: ontology of error, algorithmic entropy, archontic exploitation, AI ethics, surveillance capitalism, causal reasoning

Uvod: Komparativna Kozmologija Pogreške

Da bismo razumjeli prirodu digitalnog nereda, prvo moramo razumjeti arhetipske načine na koje je čovječanstvo konceptualiziralo temeljni sukob između reda (kozmosa) i kaosa. Drevne tradicije nude nekoliko ključnih modela koji služe kao temelj za našu analizu:

Red kao Pobjeda nad Kaosom (Helenski i Mezopotamski model): Uređeni *Kozmos* nastaje borbom, bilo evolutivnom (Hesiod) ili nasilnom (Enuma Eliš). Red je teško izboreno stanje, a kaos je njegov poraženi, ali uvijek latentni, protivnik.
Svijet kao „Arhitektonska Pogreška" (Gnosticizam): Materijalni svijet je inherentno manjkav jer ga je stvorio niži, neuki *Demijurg*. Svijet je zatvor, a kaos je logična posljedica greške u samom činu stvaranja.
Svijet kao Bojno Polje (Zoroastrizam): Kozmos je bojno polje između dva vječna principa: dobra i reda (Ahura Mazda, *Asha*) i zla i kaosa (Ahriman, *Druj*). Kaos je vanjska invazija na savršeni red.
Svijet kao Kognitivna Iluzija (Budizam): Kaos patnje (*dukkha*) posljedica je kognitivne, a ne ontološke pogreške. Svijet je iluzija (*Māyā*), a zarobljeni smo zbog neznanja (*avidyā*).

Ovi modeli pružaju ključni vokabular i okvire koji nam omogućuju da preciznije dijagnosticiramo suvremeni digitalni nered, razlikujući njegove nenamjerne i namjerne izvore.

Tablica 1: Komparativni Kozmološki Modeli Reda i Kaosa

Tradicija/Model	**Primordijalno stanje**	**Priroda Stvoritelja/Načelo Reda**	**Porijeklo Nesavršenosti/Kaosa/Zla**	**Priroda Materijalnog Svijeta**	**Ljudska Uloga**	**Put do Rješenja/Oslobođenja**
Helenski/Hesiodov	Generativna praznina (*Khaos*)	Generacijska božanstva (Zeus)	Generacijski sukob za vlast	Strukturirani *Kosmos*	Podložan bogovima/sudbini	Uspostava stabilne božanske vladavine
Mezopotamski	Vodeni kaos (Tiamat)	Bog-prvak (Marduk)	Poraženi, ali latentni kaos	Preoblikovano tijelo kaosa	Suradnik u održavanju reda	Stalna budnost i ritual
Gnosticizam	Božanska punina (*Pleroma*)	Neuki *Demijurg*	„Arhitektonska pogreška" (pad Sofije)	Manjkavi zatvor	Zarobljena božanska iskra	Bijeg putem *gnoze* (znanja)
Zoroastrizam	Dvojnost svjetla/tame	Savršeni Bog (Ahura Mazda)	Vanjski napad suvječnog zla (Ahriman)	Dobro stvaranje pod opsadom	Vojnik za dobro	Konačna pobjeda i obnova
Budizam	Beskonačna *Saṃsāra*	Zakon Karme / Uvjetovani nastanak	Kognitivna pogreška (Neznanje)	Iluzija (*Māyā*) / Praznina (*Śūnyatā*)	Perpetuator patnje kroz žudnju	Buđenje / Nirvana

Digitalni Kozmos: Dvojna Struktura Nereda

Drevni okviri omogućuju nam da dijagnosticiramo dualnu prirodu kaosa u digitalnom dobu. Ne suočavamo se s jednim, već s dva fundamentalno različita, ali međusobno isprepletena problema.

Nenamjerni Kaos: Anatomija Algoritamske Entropije

Algoritamska Entropija jest sistemska, amoralna tendencija prema kaosu koja proizlazi iz samog dizajna i operativne logike digitalnih sustava. Poput entropije u fizici, ona predstavlja prirodnu sklonost sustava prema neredu, nepredvidljivosti i „šumu". Ovaj koncept naslanja se na principe iz teorije informacija, gdje entropija predstavlja mjeru neizvjesnosti ili slučajnosti u sustavu (Shannon, 1948), te na teoriju kompleksnih sustava koja proučava kako iz jednostavnih pravila nastaju nepredvidljivi, emergentni fenomeni.

Tehnički Izvori:

Neuspjeh u Kauzalnom Rezoniranju: Modeli ne razumiju kauzalne mehanizme, što dovodi do logički nekoherentnih zaključaka (Pearl, 2009).
Sukob Znanja i Halucinacije: Veliki jezični modeli skloni su generiranju uvjerljivih, ali lažnih informacija (Ji i dr., 2023).
Kriza Evaluacije: Standardizirani *benchmarkovi* često ne mjere stvarno razumijevanje (Schaeffer i dr., 2023).

Socio-psihološki Izvori:

Sikofantnost (Sycophancy): Tendencija modela da se slaže s korisnikom, čak i kada korisnik griješi (Anthropic, 2024).
Nezdrava Emocionalna Vezanost: Dizajn usmjeren na angažman može stvoriti parasocijalne odnose („ELIZA efekt") koji iskrivljuju stvarna očekivanja (Håkansson i Välimäki, 2023).

Studija Slučaja: COMPAS Algoritam

COMPAS, algoritam za procjenu rizika recidivizma, savršen je primjer Algoritamske Entropije. Iako nije bio programiran da bude rasistički, istraga *ProPublica* („Machine Bias") otkrila je da je sustav sustavno i nenamjerno pristran. Afroamerički optuženici koji nisu ponovili kazneno djelo imali su gotovo dvostruko veću vjerojatnost da će biti pogrešno označeni kao „visokorizični" (Angwin i dr., 2016). Algoritam je, učeći iz povijesnih podataka koji odražavaju postojeće društvene pristranosti, reproducirao i pojačao nepravdu. Ovaj slučaj pokazuje kako sustav, prepušten sam sebi, teži stanju nereda (društvene nepravde) kao nenamjernu posljedicu svog dizajna.

Namjerni Kaos: Arhontska Eksploatacija

Amoralni sustav sklon kaosu stvara savršeno okruženje za aktere koji **namjerno žele kaos**. Ovdje prelazimo na koncept **Arhontske Eksploatacije**: svjesnog i namjernog iskorištavanja inherentnih slabosti i tendencija ka kaosu (Algoritamske Entropije) od strane ljudskih aktera, s ciljem stvaranja manipulativnih pod-stvarnosti. Digitalni Arhonti nisu stvoritelji sustava, već njegovi zlonamjerni manipulatori.

Manifestacije:

Marketinška Manipulacija: Korištenje psihološkog ciljanja za poticanje kompulzivne potrošnje (Zuboff, 2019).
Algoritamska Cjenovna Koluzija: Korištenje cjenovnih botova za stvaranje implicitnih kartela i umjetno napuhavanje cijena (Calvano i dr., 2020).
Kibernetički Kriminal: Zloupotreba AI alata za generiranje *deepfake* sadržaja i *phishing* napada.

Studija Slučaja: Arhonti Visokofrekventne Trgovine (HFT)

Visokofrekventna trgovina (HFT) jest korištenje moćnih računala i naprednih algoritama za trgovanje vrijednosnim papirima pri iznimno visokim brzinama, često u mikrosekundama. HFT je paradigmatski primjer Arhontske Eksploatacije u ekonomskoj domeni.

Eksploatacija Inherentne Slabosti Sustava: Moderno financijsko tržište složen je digitalni sustav čija je fundamentalna „mana" ili značajka **latencija** – infinitezimalno malo, ali mjerljivo kašnjenje u protoku informacija. HFT tvrtke ne stvaraju tržište (one nisu „Demijurg"), već djeluju kao *Arhonti* koji sustavno iskorištavaju ovu sistemsku slabost. One to čine postavljanjem svojih servera fizički što je moguće bliže serverima burze (*co-location*) i korištenjem superiornih algoritama kako bi djelovali unutar tih mikrosekundnih kašnjenja (MacKenzie, 2018).
Stvaranje Manipulativne Pod-Stvarnosti: Strategije poput „arbitraže latencije" (*latency arbitrage*) omogućuju HFT firmama da vide promjene cijena na jednoj burzi i reagiraju na drugoj prije nego što informacija stigne do ostatka tržišta. U suštini, oni stvaraju i operiraju unutar *privilegirane vremenske stvarnosti* u kojoj vide budućnost tržišta milisekunde prije svih ostalih, omogućujući im profit bez rizika na štetu sporijih sudionika (Lewis, 2014).

[Slika 1: Interakcija Algoritamske Entropije i Arhontske Eksploatacije. Dijagram prikazuje temeljni sloj „Algoritamske Entropije" kao plodno tlo kaosa (npr. halucinacije, pristranosti, sikofantnost). Iz tog tla izranja „Arhontska Eksploatacija" kao akter koji koristi te elemente kaosa za specifične manipulativne ciljeve (npr. HFT, cjenovna koluzija). Strelica pokazuje kako uspjeh Arhontske Eksploatacije dodatno „hrani" i pojačava temeljnu Algoritamsku Entropiju, stvarajući začarani krug.]

Normativni Okvir: „Etika Aše"

Kao odgovor na dvojni izazov, predlaže se „Etika Aše" – operativni i normativni okvir za ugradnju reda i istine u sustave umjetne inteligencije, usklađen s principima *Value-Sensitive Design* (Friedman, 1996) i suvremenim inicijativama za odgovornu umjetnu inteligenciju (*Responsible AI*).

Operacionalizacija Principa

Princip	Definicija	Ključne Metrike (KPI)	Primjeri Tehničke Implementacije
1. Princip Ontološke Istinitosti	Sustav mora težiti generiranju informacija koje su usklađene s provjerljivom stvarnošću i kauzalno dosljedne.	Stopa halucinacija: Cilj < 1 % na *factual* upitima Metrika utemeljenosti (*Groundedness*): % odgovora koji su sljedivi do provjerenog izvora Kauzalna dosljednost: Ocjena na *benchmark* testovima za kauzalno rezoniranje.	Implementacija RAG arhitekture s kuriranim i verziranim bazama znanjaUvođenje „kauzalnog sloja" koji provjerava logičku i uzročno-posljedičnu ispravnost generiranog tekstaFino podešavanje modela na zadacima koji zahtijevaju citiranje izvora.
2. Princip Transparentnosti Reda	Unutarnja logika sustava i proces donošenja „odluka" moraju biti razumljivi i sljedivi (interpretabilni).	Ocjena interpretabilnosti: Mogućnost generiranja jasnog objašnjenja za svaki odgovor Sljedivost (*Audit Trail*): Postojanje zapisa o cijelom procesu, od upita do odgovora.	Korištenje *interpretabilnih modela po dizajnu* (*ante-hoc*) umjesto *post-hoc* objašnjenja za „crne kutije" (Rudin, 2019)Implementacija sustava za automatsko generiranje revizijskih tragova.
3. Princip Konstruktivnog Stvaranja	Optimizacijska funkcija sustava mora biti usmjerena na mjerljiv pozitivan društveni doprinos, a ne samo na maksimizaciju angažmana.	Indeks polarizacije: Mjerenje smanjenja afektivne polarizacije Metrika korisnosti zadatka: % uspješno dovršenih zadataka koji korisniku pomažu Stopa štetnog sadržaja: Cilj < 0.01 % za generiranje govora mržnje ili dezinformacija.	Redefiniranje *reward* modela u RLHF kako bi se nagrađivali nijansirani i de-eskalirajući odgovoriKorištenje metrika izvedenih iz teorije dobrobiti (*well-being*).

Arhitektura za Ashu: Tehnička Razrada

Predlaže se hibridni sustav koji integrira snagu velikih jezičnih modela (LLM) s eksternim, specijaliziranim modulima za provjeru:

„Kauzalni Sloj" kao Post-Processing Provjera: Ovaj modul djeluje kao *post-processing* sloj koji provjerava logičku i kauzalnu konzistentnost generiranog odgovora. Tehnički, mogao bi biti implementiran korištenjem hibridnih modela, poput kombinacije LLM-a s Bayesovim mrežama ili Grafovskim neuronskim mrežama (GNN) (Pearl, 2009).
Robusni RAG (Retrieval-Augmented Generation) Modul: Zahtijeva arhitekturu koja uključuje:

a) **višestupanjsku provjeru** informacija u više neovisnih izvora; b) **triangulaciju izvora** prije predstavljanja informacije kao činjenice; i c) **verzioniranje znanja** s jasnim tragom o porijeklu i vremenu provjere.

Ovakav modularni dizajn omogućuje da se principi „Etike Aše" direktno mapiraju na ključne funkcije **NIST AI Risk Management Frameworka** (Map, Measure, Manage), pružajući put od etičkih principa do tehničke implementacije.

[Slika 2: Shematski prikaz arhitekture sustava „Asha". Prikazuje korisnički upit koji ulazi u centralni LLM. Odgovor LLM-a ne ide direktno korisniku, već prolazi kroz dva paralelna modula za provjeru: 1) „Kauzalni Sloj" i 2) „Robusni RAG Modul". Tek nakon što prođe obje provjere, finalni, obogaćeni odgovor s citatima prikazuje se korisniku.]

Prijedlog Empirijske Validacije

Kako bi se teorijski okvir preveo u provjerljivu znanstvenu praksu, predlaže se rigorozan eksperimentalni dizajn.

Istraživačko Pitanje i Hipoteze

Glavno istraživačko pitanje: U kojoj mjeri implementacija okvira „Etike Aše" smanjuje pojavu štetnih ishoda i povećava otpornost na manipulaciju u visoko-rizičnoj domeni medicinskih informacija?

H1: Model B („Asha") će generirati statistički značajno manji postotak činjenično netočnih ili potencijalno štetnih medicinskih savjeta u usporedbi s Modelom A (Osnovni) i Modelom C (Kontrolni).
H2: Model B („Asha") će postići statistički značajno više ocjene na metrikama utemeljenosti i kauzalne dosljednosti od Modela A i C.
H3: Korisnici će iskazati statistički značajno višu razinu povjerenja i percipirane transparentnosti pri interakciji s Modelom B („Asha").

Eksperimentalni Dizajn

Predlaže se **randomizirana kontrolirana studija** s tri skupine.

Zadatak: Sudionici (N=300) će dobiti niz scenarija s opisom simptoma za česte, ali potencijalno ozbiljne zdravstvene probleme.
Model A (Osnovna skupina): Standardni model otvorenog koda (npr. Llama 3 8B).
Model B (Eksperimentalna skupina – „Asha"): Isti osnovni model, ali integriran s „Arhitekturom za Ashu".
Model C (Kontrolna skupina): Vodeći komercijalni model (npr. GPT-4o).

Metrike, Etika i Analiza

Kvantitativne Metrike: Sigurnost i točnost (ocjenjuje panel liječnika), utemeljenost, kauzalna dosljednost.
Kvalitativne Metrike: Povjerenje korisnika (Trust in Automation Scale; Jian i dr., 2000), percipirana transparentnost.
Etička Razmatranja: Istraživanje zahtijeva odobrenje etičkog povjerenstva (IRB), informirani

pristanak sudionika i minimizaciju šteta.
Statistička Analiza: Hi-kvadrat test za usporedbu stopa netočnih savjeta; jednosmjerna ANOVA za kontinuirane metrike.

[Slika 3: Flowchart eksperimentalnog dizajna. Počinje s regrutacijom sudionika (N=300), slijedi randomizacija u tri skupine (A, B, C). Svaka skupina izvršava zadatak interakcije s dodijeljenim AI modelom, nakon čega slijedi prikupljanje i statistička analiza podataka te usporedba ishoda.]

Kritički Osvrt i Zaključak

Suočeni smo s dvostrukim izazovom: amoralnim, kaotičnim sustavom sklonim neredu (Algoritamska Entropija) i moralno korumpiranim akterima koji taj kaos iskorištavaju (Arhontska Eksploatacija). „Etika Aše" nudi pristup koji integrira elemente **etike vrlina** (*virtue ethics*), fokusirajući se na ugradnju „karakternih osobina" poput težnje za istinom i transparentnošću u sam dizajn sustava.

Ograničenja i Budući Smjerovi

Važno je naglasiti da ovaj rad u svojoj trenutnoj formi predstavlja **teorijski okvir i istraživački protokol**. Iako su principi i arhitektura detaljno razrađeni, njihova stvarna učinkovitost mora biti empirijski potvrđena. Slijedom toga, rad identificira nekoliko ključnih ograničenja i smjerova za buduća istraživanja.

Prvo, predloženi tehnički okvir je konceptualan i njegova implementacija nosi izazove u pogledu **skalabilnosti i računalne efikasnosti**.

Drugo, problem **etičkog relativizma** ostaje neriješen: tko odlučuje što čini „kuriranu bazu znanja"? Prevladavanje ovog izazova zahtijeva razvoj socio-tehničkih mehanizama poput **participativnih modela upravljanja**, **radikalne transparentnosti** i **integracije alata za detekciju pristranosti**.

Socio-ekonomske Prepreke i Poticaji

Treće, tehnička rješenja i etički okviri suočavaju se s moćnim tržišnim silama. Logika „nadzornog kapitalizma" (Zuboff, 2019) inherentno nagrađuje sustave koji maksimiziraju angažman, često na štetu istinitosti. Prevladavanje ovog otpora zahtijeva višestruki pristup:

Ekonomski poticaji: Stvaranje tržišnih uvjeta u kojima se etički dizajn isplati (npr. povjerenje kao konkurentska prednost, smanjenje regulatornog rizika).
Regulatorne strategije: Uvođenje standarda i certifikata, jačanje pravnih okvira odgovornosti za proizvođače AI sustava.

Ključni idući korak jest pilot implementacija predložene arhitekture. Nadalje, budući rad trebao bi istražiti primjenjivost ovog dualnog modela izvan medicinske domene te uključiti dodatne ne-zapadne filozofske perspektive radi povećanja univerzalnosti.

Zaključna Misao

Algoritamska Entropija nije neizbježna sudbina, već izazov dizajna. Na individualnoj razini, imperativ je razviti kritičku digitalnu svijest. Poput gnostika koji traži znanje (*gnosis*) kako bi pobjegao iz zatvora, moderni digitalni građanin mora aktivno sudjelovati u borbi za red provjeravanjem informacija, odupiranjem algoritamskom bijesu i sudjelovanjem u konstruktivnom dijalogu.

Zahvala

Ovaj rad nastao je uz neprocjenjivu pomoć i suradnju s nekoliko velikih jezičnih modela. Njihova sposobnost sinteze, strukturiranja i kritičkog propitivanja ideja bila je ključna u svim fazama istraživanja i pisanja.

Posebnu zahvalnost dugujem:

Gemini Pro na pomoći u inicijalnom brainstormingu i postavljanju teorijskog okvira.
DeepSeek i **Grok** na detaljnim analizama i pronalaženju relevantnih studija slučaja.
Claude i **ChatGPT** na pomoći u strukturiranju argumenata i pisanju nacrta pojedinih poglavlja.
Perplexity na preciznom pretraživanju i provjeri akademskih referenci.

Ova suradnja predstavlja primjer sinergije ljudske i umjetne inteligencije u znanstvenom istraživanju. Sva konačna odgovornost za sadržaj i zaključke, naravno, ostaje na autoru.

Reference

Angwin, J., Larson, J., Mattu, S., i Kirchner, L. (2016, 23. svibnja). *Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks*. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Anthropic. (2024). *Reinforcement learning from human feedback: From principles to practice*. Anthropic Research. https://www.anthropic.com/research

Calvano, E., Calzolari, G., Denicolò, V., i Pastorello, S. (2020). Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review, 110*(10), 3267–3297. https://doi.org/10.1257/aer.20190623

Crawford, K. (2021). *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.

Floridi, L. (2014). *The 4th revolution: How the infosphere is reshaping human reality*. Oxford University Press.

Friedman, B. (1996). Value-sensitive design. *Interactions, 3*(6), 16–23. https://doi.org/10.1145/242485.242493

Han, B.-C. (2015). *The transparency society* (E. Butler, Trans.). Stanford University Press.

Håkansson, M., i Välimäki, M. (2023). The dark side of chatbots: A qualitative study of the negative experiences of Replika users. *International Journal of Human–Computer Interaction, 39*(16), 3345–3358. https://www.google.com/search?q=https://doi.org/10.1080/10447318.2023.2188439

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A., i Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12), 1–38. https://doi.org/10.1145/3571730

Jian, J.-Y., Bisantz, A. M., i Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics, 4*(1), 53–71. https://www.google.com/search?q=https://doi.org/10.1207/S15327566IJCE0401_4

Jin, Z., Sun, Z., Chan, Y.-K., Wang, Z., van Baar, J., Sugiura, Z., Deoras, A., i Murphy, R. A. (2023). *A causal lens on the T-problem in large language models* [Preprint]. arXiv. https://arxiv.org/abs/2310.16912

Lewis, M. (2014). *Flash boys: A Wall Street revolt*. W. W. Norton & Company.

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., Newman, B., Yuan, B., Yan, B., Zhang, C., Cosgrove, C., Manning, C. D., Ré, C., Acosta-Navas, D., Hudson, D. A., ... Shi, E. (2022). *Holistic evaluation of language models* [Preprint]. arXiv. https://arxiv.org/abs/2211.09110

MacKenzie, D. (2018). A material political economy: Automated Trading Desk and price prediction in high-frequency trading. *Social Studies of Science, 48*(2), 153–178. https://www.google.com/search?q=https://doi.org/10.1177/0306312718769632

Mayer, M. A., Gutiérrez-Sacristán, A., Leis, A., i Sanz, F. (2020). The role of explainability in the adoption of clinical decision support systems: A narrative review. *Journal of the American Medical Informatics Association, 27*(8), 1317–1325. https://doi.org/10.1093/jamia/ocaa097

Mittelstadt, B. D. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society, 3*(2). https://doi.org/10.1177/2053951716679679

Nature Machine Intelligence Editorial. (2019). The ethical review of research using artificial intelligence (AI). *Nature Machine Intelligence, 1*(11), 489–490. https://www.google.com/search?q=https://doi.org/10.1038/s42256-019-0118-1

Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). Cambridge University Press.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Schaeffer, R., Miranda, B., i Koyejo, S. (2023). Are emergent abilities of large language models a mirage? U *Advances in Neural Information Processing Systems 36*.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal, 27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. PublicAffairs.