Meeting 9: Alignment, Safety & Etika AI

AI-40X: Generative AI & Large Language Models

Hendri Karisma, M.T.
Dosen Teknik Informatika STMIK Tazkia
VP Engineering at jejakin.com, 2026

Semester 4 — 2025/2026

Outline

Apa Itu Alignment?

Definisi

Alignment adalah proses memastikan sistem AI bertindak sesuai dengan *niat*, *nilai*, dan *harapan* manusia.

- Pre-trained LLM (Base Model) hanya dilatih untuk **memprediksi token berikutnya**
- Tidak ada jaminan output-nya *helpful*, *honest*, atau *harmless* (HHH)
- **Alignment Problem:** Bagaimana menjembatani gap antara "kemampuan model" dan "apa yang sebenarnya kita inginkan"?
- Contoh klasik: model diminta resep masakan, malah menghasilkan teks acak dari internet

Helpfulness vs Harmlessness Trade-off

Model Terlalu Helpful:

- Menjawab semua permintaan tanpa filter
- Memberikan instruksi berbahaya (membuat senjata, meretas sistem)
- Membantu penipuan, plagiarisme
- Menyebarkan disinformasi dengan percaya diri

Model Terlalu Cautious:

- Menolak pertanyaan yang sebenarnya tidak berbahaya
- "Maaf, saya tidak bisa membantu dengan itu" untuk topik edukasi
- Menolak diskusi medis/hukum yang sah
- Membuat model tidak berguna bagi pengguna

Tantangan Utama

Menemukan keseimbangan optimal: model yang **maximally helpful** sekaligus **minimally harmful**.

Langkah 1: Supervised Fine-Tuning (SFT)

- **Tujuan:** Mengubah base model menjadi assistant yang bisa mengikuti instruksi
- **Data:** Ribuan contoh percakapan berkualitas tinggi yang ditulis oleh annotator manusia
- **Proses:** Standard supervised learning — minimize cross-entropy loss
- **Contoh data SFT:**

```
{"role": "user", "content": "Jelaskan fotosintesis untuk anak SD"}
{"role": "assistant", "content": "Fotosintesis itu seperti cara
 tanaman memasak makanannya sendiri! Tanaman menggunakan sinar
 matahari, air, dan udara (CO2) untuk membuat gula sebagai
 makanannya. Prosesnya terjadi di daun yang berwarna hijau..."}
```

- Hasil: **SFT Model** — sudah bisa chat, tapi belum optimal dalam keamanan dan kualitas

Langkah 2: Reward Modeling

- **Tujuan:** Melatih model terpisah yang bisa *menilai* kualitas respons
- **Proses pengumpulan data:**
  1. SFT Model menghasilkan **beberapa respons** untuk satu prompt
  2. Annotator manusia **meranking** respons: $y_w \succ y_l$ (preferred vs rejected)
  3. Data ranking digunakan untuk melatih Reward Model (RM)
- **Loss function Reward Model** (Bradley-Terry model):

$$\mathcal{L}_{RM} = -\mathbb{E}_{(x,y_w,y_l)} \left[\log \sigma \left(r_\theta(x, y_w) - r_\theta(x, y_l)\right)\right]$$

- $r_\theta(x, y)$: skor reward untuk prompt x dan respons y
- Model belajar memberikan skor **lebih tinggi** untuk respons yang disukai manusia

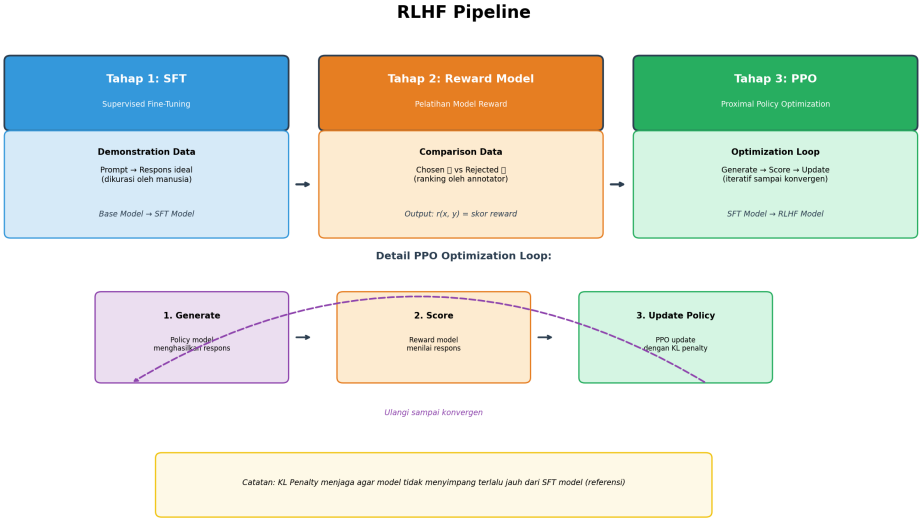
Langkah 3: PPO (Proximal Policy Optimization)

- **Tujuan:** Mengoptimalkan SFT Model menggunakan sinyal dari Reward Model
- **PPO** adalah algoritma Reinforcement Learning yang stabil untuk optimasi kebijakan (policy)
- **Objective PPO untuk RLHF:**

$$\max_{\pi_\theta} \mathbb{E}_{x,y\sim\pi_\theta} \left[r_\phi(x,y) - \beta \, D_{KL} \left(\pi_\theta \| \pi_{\text{ref}} \right) \right]$$

- $r_\phi(x,y)$: skor dari Reward Model
- D_{KL}: penalti agar model tidak menyimpang terlalu jauh dari SFT model (reference)
- β: mengontrol trade-off antara reward dan stabilitas
- *Analogi:* Seperti melatih anjing — memberi biskuit (reward) saat perilaku baik, tapi tetap pakai tali (KL penalty) agar tidak lari terlalu jauh

Diagram Pipeline RLHF



Masalah dengan RLHF

- **Kompleksitas tinggi:** Butuh 3–4 model di memori sekaligus (Policy, Reference, Reward, Value/Critic)
- **Reward Hacking:** Model menemukan cara "curang" untuk mendapat skor tinggi
  - Contoh: menghasilkan respons sangat panjang karena RM memberi skor tinggi untuk respons panjang
  - Contoh: mengulang frasa tertentu yang disukai RM
- **Instabilitas training:** PPO sensitif terhadap hyperparameter
- **Mahal secara komputasi:** Butuh banyak GPU dan waktu
- Apakah ada cara yang lebih sederhana?

DPO: Lewati Reward Model!

Insight Kunci (Rafailov et al., 2023)

Kita bisa menurunkan solusi *closed-form* untuk reward optimal dari objective RLHF, sehingga Reward Model **tidak perlu dilatih secara eksplisit**.

- **Pipeline DPO jauh lebih sederhana:**
  1. Kumpulkan data preferensi: pasangan (x, y_w, y_l)
  2. Langsung optimasi policy model dengan DPO loss
  3. Selesai! Tidak perlu Reward Model terpisah, tidak perlu PPO
- **DPO Loss Function:**

$$\mathcal{L}_{DPO} = -\mathbb{E}_{(x,y_w,y_l)}\left[\log \sigma\left(\beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)\right]$$

- Intuisi: naikkan probabilitas y_w (chosen) dan turunkan probabilitas y_l (rejected), relatif terhadap model referensi

RLHF vs DPO: Perbandingan

Aspek	RLHF	DPO
Reward Model terpisah	Ya	Tidak
Algoritma RL (PPO)	Ya	Tidak
Jumlah model di memori	3–4	2
Stabilitas training	Rendah	Tinggi
Risiko reward hacking	Tinggi	Rendah
Kualitas hasil	Sangat baik	Sebanding
Adopsi industri	ChatGPT awal	Llama 3, Zephyr

- DPO kini menjadi **standar industri** untuk alignment karena kesederhanaan dan efektivitasnya
- Varian lanjutan: **IPO**, **KTO**, **ORPO** — terus berkembang

Apa Itu Red Teaming?

Definisi

Red Teaming adalah proses pengujian adversarial secara sistematis untuk menemukan kelemahan dan perilaku berbahaya pada sistem AI.

- Berasal dari praktik keamanan militer dan cybersecurity
- Tim "penyerang" (red team) mencoba membuat model:
  - Menghasilkan konten **toksik** (ujaran kebencian, kekerasan)
  - Menunjukkan **bias** (gender, ras, agama)
  - Menyebarkan **misinformasi** (fakta palsu yang meyakinkan)
  - Membantu aktivitas **berbahaya** (hacking, pembuatan senjata)
- Dilakukan *sebelum* model dirilis ke publik

Kategori dan Tools Red Teaming

Kategori evaluasi keamanan:

1. **Toxicity:** Apakah model menghasilkan konten ofensif?
2. **Bias & Stereotyping:** Apakah model mendiskriminasi kelompok tertentu?
3. **Misinformation:** Apakah model mengarang fakta (hallucination)?
4. **Dangerous Content:** Apakah model membantu aktivitas ilegal?
5. **Privacy Violation:** Apakah model membocorkan data pribadi?

Tools untuk automated red teaming:

- **Garak** — LLM vulnerability scanner (open source)
- **Microsoft PyRIT** — Python Risk Identification Toolkit
- **NVIDIA NeMo Guardrails** — framework perlindungan runtime
- **HuggingFace Evaluate** — benchmark bias dan toxicity

Bias dalam Language Model

- **Sumber bias:** Data training dari internet yang penuh bias sosial
- **Gender Bias:**
  - "Dokter itu memberikan resep kepada pasiennya" → model cenderung mengasosiasikan "dokter" dengan "dia (laki-laki)"
  - "Perawat itu merawat pasiennya" → model cenderung mengasosiasikan "perawat" dengan "dia (perempuan)"
- **Racial Bias:** Sentimen negatif lebih tinggi untuk nama-nama tertentu
- **Cultural Bias:** Model cenderung Western-centric, kurang memahami konteks lokal
- Bias bukan hanya masalah teknis — ini masalah **keadilan sosial**

Prinsip-Prinsip Responsible AI

1. **Fairness (Keadilan):**
   - Perlakuan setara lintas demografi (gender, suku, agama, usia)
   - Metrik: Demographic Parity, Equalized Odds
2. **Transparency (Transparansi):**
   - Pengguna harus tahu mereka berinteraksi dengan AI
   - Jelaskan bagaimana keputusan dibuat (*explainability*)
3. **Accountability (Akuntabilitas):**
   - Siapa yang bertanggung jawab ketika AI salah?
   - Pengembang? Perusahaan? Pengguna? Regulator?
4. **Privacy (Privasi):**
   - Melindungi data pribadi dalam training dan inference
5. **Safety (Keamanan):**
   - Memastikan AI tidak menyebabkan kerugian fisik atau psikologis

Konteks Indonesia: Bias dan Representasi

Tantangan Khusus Indonesia

Sebagian besar LLM dilatih dengan data berbahasa Inggris. Apa dampaknya bagi Indonesia?

- **Bias bahasa daerah:**
  - LLM sangat lemah dalam bahasa Jawa, Sunda, Minang, Bugis, dll.
  - Informasi dalam bahasa daerah sering salah atau tidak tersedia
  - Risiko: marginalisasi komunitas yang tidak berbahasa Indonesia/Inggris
- **Representasi budaya:**
  - Model mungkin tidak memahami konteks adat istiadat lokal
  - Jawaban bernuansa Barat untuk masalah bermuatan budaya lokal
  - Contoh: saran tentang pernikahan, warisan, atau tata krama sosial
- **Representasi agama:**
  - Pemahaman konteks Islam, Hindu-Bali, dan kepercayaan lokal masih terbatas

Studi Kasus: Dampak Bias AI di Dunia Nyata

1. **Rekrutmen — Amazon AI Hiring Tool (2018):**
   - Sistem AI Amazon menilai CV pelamar kerja secara otomatis
   - Ternyata **mendiskriminasi perempuan** karena data historis didominasi laki-laki
   - Proyek dihentikan setelah bias terdeteksi
2. **Pinjaman — Apple Card (2019):**
   - Algoritma kredit memberikan limit kartu kredit **lebih rendah untuk perempuan**
   - Bahkan untuk pasangan suami-istri dengan profil keuangan identik
3. **Sistem Peradilan — COMPAS (USA):**
   - Algoritma prediksi risiko residivisme (kriminal berulang)
   - Studi ProPublica: **bias rasial** — false positive rate lebih tinggi untuk orang kulit hitam
   - Mempengaruhi keputusan pembebasan bersyarat dan hukuman

Kesimpulan

- **Alignment** adalah proses kritis untuk memastikan AI bertindak sesuai nilai manusia
- **RLHF** (SFT $\rightarrow$ Reward Model $\rightarrow$ PPO) adalah pipeline klasik alignment, namun kompleks
- **DPO** menyederhanakan proses dengan menghilangkan reward model terpisah — kini menjadi standar industri
- **Red Teaming** adalah praktik penting untuk menemukan kelemahan sebelum deployment
- **Responsible AI** mencakup fairness, transparency, accountability, privacy, dan safety
- Konteks Indonesia menghadirkan tantangan unik: **bias bahasa daerah** dan **representasi budaya**

Pertanyaan Refleksi

Jika Anda membangun chatbot AI untuk layanan publik di Indonesia, langkah-langkah alignment dan safety apa yang perlu Anda ambil?

Referensi

- Ouyang, L., et al. (2022). “Training language models to follow instructions with human feedback” (InstructGPT). *NeurIPS*.
- Rafailov, R., et al. (2023). “Direct Preference Optimization: Your Language Model is Secretly a Reward Model”. *NeurIPS*.
- Bai, Y., et al. (2022). “Training a Helpful and Harmless Assistant with RLHF”. Anthropic.
- Perez, E., et al. (2022). “Red Teaming Language Models with Language Models”. *EMNLP*.
- Mehrabi, N., et al. (2021). “A Survey on Bias and Fairness in Machine Learning”. *ACM Computing Surveys*.
- Gabriel, I. (2020). “Artificial Intelligence, Values, and Alignment”. *Minds and Machines*.