

Crash Course Ethical & Trustworthy AI

6th August 2023

Ilma Aliya Fiddien
ilmaaliyaf@gmail.com

In collaboration with



Indonesia AI

AI for Everyone, AI for Indonesia

Outline

1. Ethical AI
2. Case study 1: Magic Avatar
3. Trustworthy AI
4. Case study 2: Bolo/Read Along

Outcomes

Learning Ethical AI

1. Ethical appreciation/sensitivity
2. Ethical analysis
3. Ethical decision-making/
applied ethics

... in the realm of AI

Outcomes

Learning Trustworthy AI

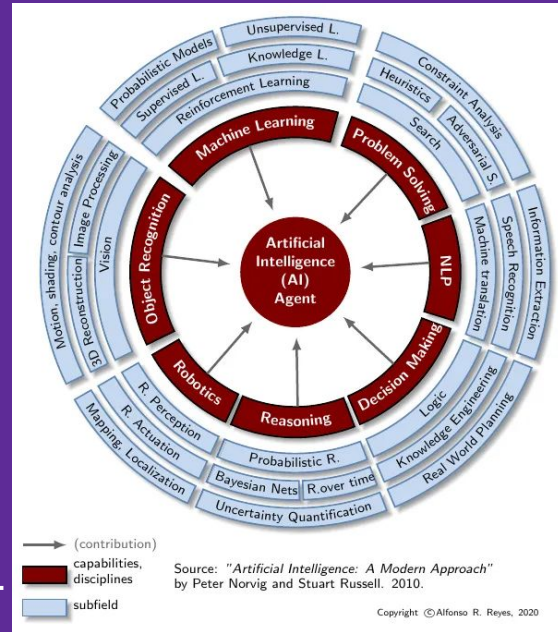
1. Awareness of existing frameworks of Trustworthy AI
 2. General understanding of applying the frameworks
-

Ethical AI

pengembangan dan pengaplikasian
teknologi kecerdasan artifisial
yang sesuai dengan nilai-nilai etika tertentu

Mengapa peduli?

artificial intelligence
≠
machine learning



Mengapa peduli?

Artificial intelligence

“[The automation of] activities that we associate with human thinking, activities such as decision-making, problem-solving, learning...”

Bellman, R. E. (1978). *An introduction to artificial intelligence: Can computers think?* San Francisco, CA: Boyd & Fraser Publishing Company.

Mengapa peduli?

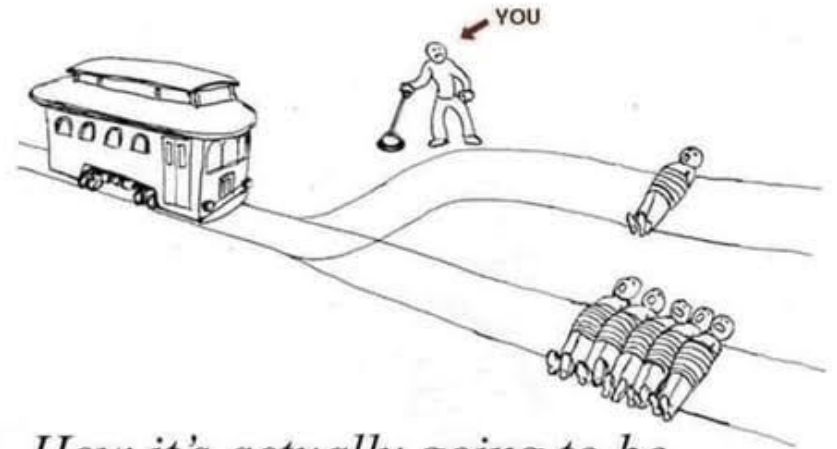
Di setiap teknologi...

- Ada **keuntungan/kerugian**
 - Selalu relevan
 - Ada **etika**
 - Relevan ketika ada keputusan/aksi yang berimplikasi moral
 - Ada **regulasi**
 - Relevan ketika ada kekhawatiran yang signifikan di terhadap individu/masyarakat/lingkungan
-

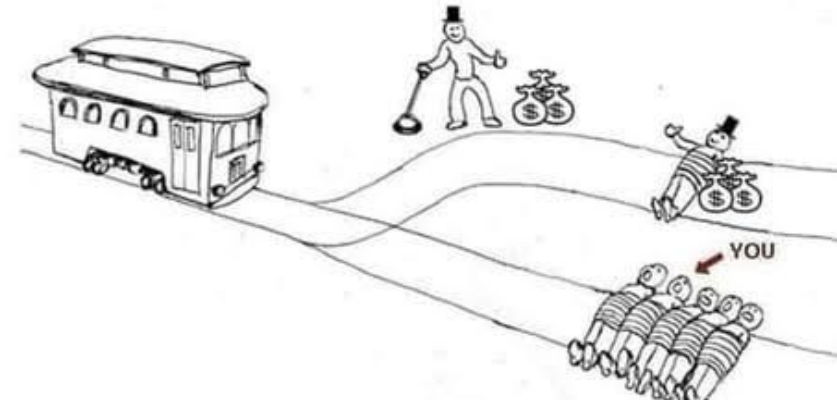
Mengapa peduli?



How you imagine the trolley problem



How it's actually going to be



Keuntungan/ Kerugian

- Kebutuhan
- Preferensi
- Privasi
- Harga diri
- Kepercayaan (trust)



Implikasi moral,
keuntungan/kerugian yang
signifikan secara etis

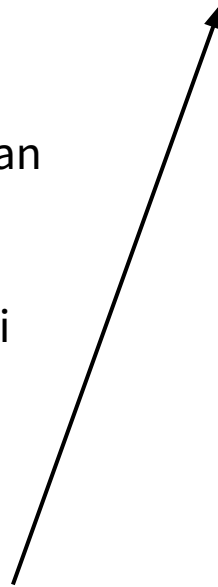


Etika

- Agency (pemanfaatan kekuasaan)
- Pilihan dan konsekuensi
- Nilai dan keyakinan



Kekhawatiran
yang signifikan



Regulasi

- Panduan etika
- Akuntabilitas
- Audit dan kepatuhan
- Pendidikan
- Tenaga kerja
- Funding

Contoh: Big Data

Keuntungan/ Kerugian

Perusahaan:
keputusan berbasis data

Konsumen:
pelanggaran privasi



Implikasi moral,
keuntungan/kerugian yang signifikan secara etis

Etika

Kepemilikan:
konsumen bisa meminta datanya dihapus
Izin/consent:
perusahaan harus meminta izin konsumen
Privasi: setiap data terjaga kerahasiaannya



Kekhawatiran yang signifikan

Regulasi



UNDANG-UNDANG REPUBLIK INDONESIA
NOMOR 27 TAHUN 2022
TENTANG
PELINDUNGAN DATA PRIBADI

Etika

berfokus pada mendefinisikan dan mewujudkan “hidup yang baik”

**Suatu keuntungan/kerugian
dipandang signifikan secara etis**

**ketika ada kemungkinan besar ia memengaruhi
kemampuan seseorang/sekelompok orang
untuk bisa hidup dengan baik.**

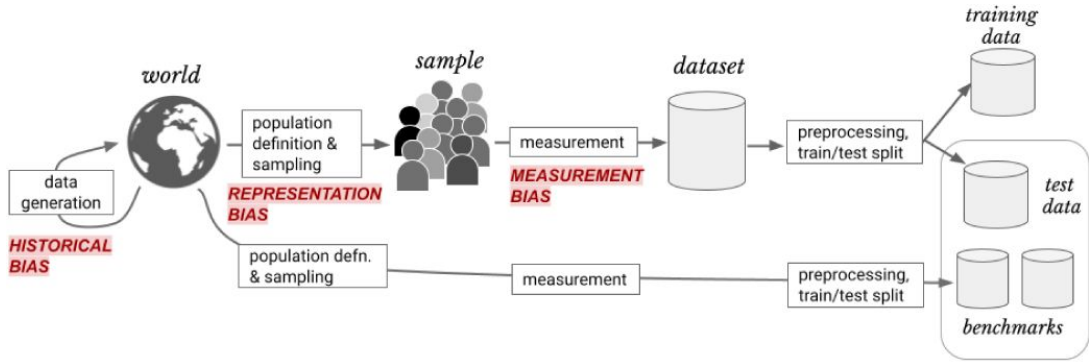
Siapa yang menerima keuntungan/kerugian?

Hampir semua orang.

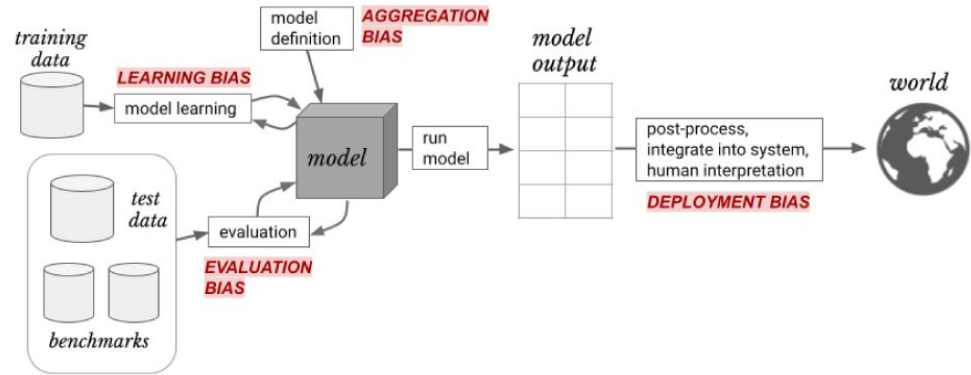
Sistem AI telah digunakan untuk menjawab pertanyaan:

- Siapa yang diterima?
- Siapa yang dipekerjakan?
- Siapa yang dipromosikan?
- Siapa yang menerima pinjaman?
- Siapa yang menerima perawatan untuk masalah medis?
- Siapa yang menerima hukuman mati?

Dari bagian mana saja dari proses pengembangan model AI yang berpengaruh?



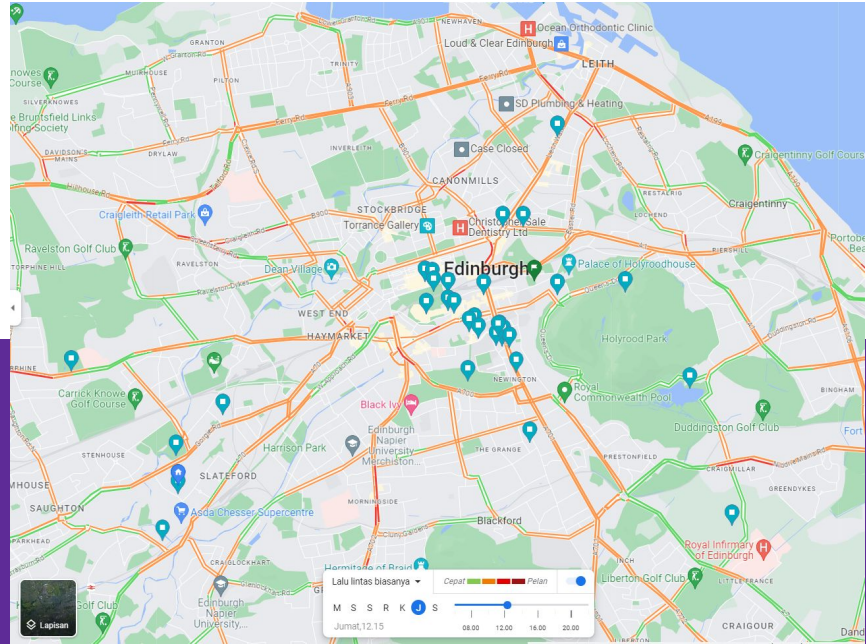
(a) Data Generation



(b) Model Building and Implementation

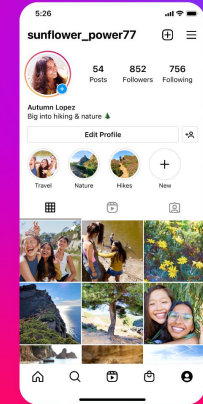
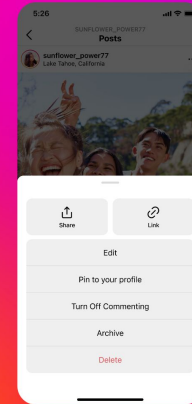
Manfaat?

Mudarat?



Manfaat?

Mudarat?




Manfaat?

Mudarat?



Lensa AI: photo editor, video 4+
Magic avatars, picture effects
[Prisma labs, inc.](#)
#74 in Photo & Video
★★★★★ 4.5 • 3.7K Ratings
Free · Offers In-App Purchases



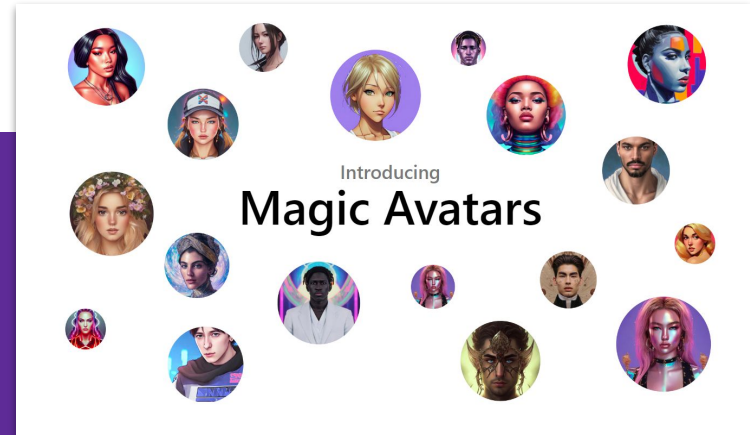
Introducing
Magic Avatars

Pengguna: keseruan

Penyedia: keuntungan bisnis
(pemasukan USD 3.99 untuk
pengeluaran USD 0.50)

Pengguna: bias fitur wajah,
penambahan elemen seksual yang
tidak perlu

Pelaku seni: persaingan



[Lensa AI climbs the App Store charts as its 'magic avatars' go viral | TechCrunch](#)
[A rocky past haunts the mysterious company behind the Lensa AI photo app : NPR](#)
[Users Complain That Lensa AI Selfie Generator is 'Sexualizing' Their Photos | PetaPixel](#)
[Lensa, the AI portrait app, has soared in popularity. But many artists question the ethics of AI art.](#)

Area Diskusi Etika/Filsafat Moral

1. Meta-etika

- Etika mengenai etika
- Apa maksud dari “benar” dan “salah”?

2. Etika preskriptif (*normative ethics*)

- Kriteria “benar” dan “salah”
- 3 kategori besar:
 - i. *Deontological ethics*
 - ii. *Teleological ethics*
 - iii. *Virtue ethics*

3. Etika deskriptif (*comparative ethics*)

- Studi empiris tentang etika yang ada di suatu lingkungan

4. Etika aplikatif (*applied ethics*)

- Pengaplikasian etika preskriptif untuk masalah praktis
- Contoh:
 - i. *Medical ethics*
 - ii. *Bioethics*
 - iii. *Environmental ethics*
 - iv. *AI ethics*

Suatu aksi itu “benar” jika ia...



Deontological ethics

Fokus: mengikuti aturan

- sesuai dengan **aturan** moral
- aksi tersebut bagus secara inheren



Teleological ethics

Fokus: memaksimalkan hasil

- menghasilkan **konsekuensi** terbaik
- aksi tersebut hanyalah instrumen mencapai akhir

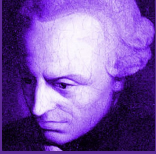


Virtue ethics

Fokus: memelihara karakter

- sesuai dengan apa yang akan dilakukan oleh orang **bijak**/saleh di situasi tersebut

Dipandu oleh nilai...



Deontological ethics

Fokus: mengikuti aturan

kebenaran | *right*
(*rationality is doing one's moral duty*)



Teleological ethics

Fokus: memaksimalkan hasil

kebaikan | *good*
(*maximum utility*)



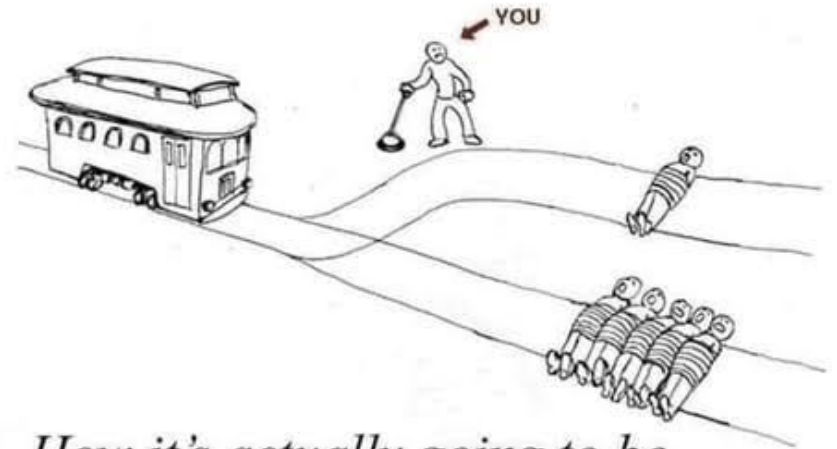
Virtue ethics

Fokus: memelihara karakter

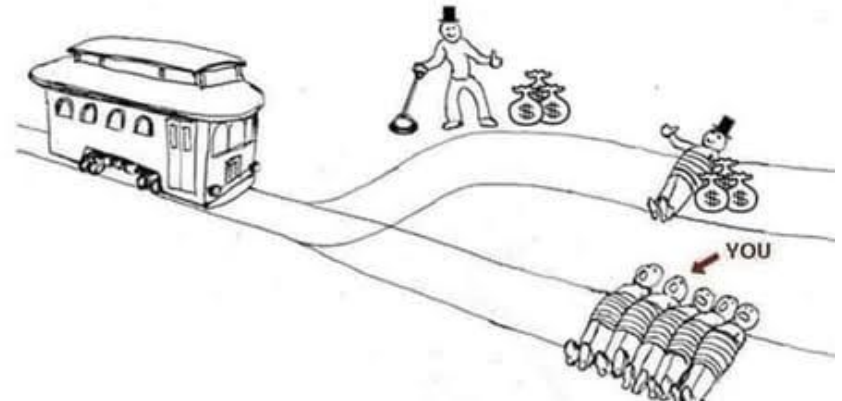
kebajikan | *virtue*
(*leading to attainment of eudaimonia/harmony*)

The trolley problem

How you imagine the trolley problem



How it's actually going to be



Respon untuk *the trolley problem*...



Deontological ethics

Fokus: mengikuti aturan

menyakiti orang tak bersalah secara aktif adalah salah secara moral, terlepas dari potensi kebaikan yang lebih besar → jangan pegang kendali tuas



Teleological ethics

Fokus: memaksimalkan hasil

minimalisasi kerugian → lima nyawa lebih berharga dari satu nyawa → mengarahkan troli ke trek dengan lebih sedikit orang



Virtue ethics

Fokus: memelihara karakter

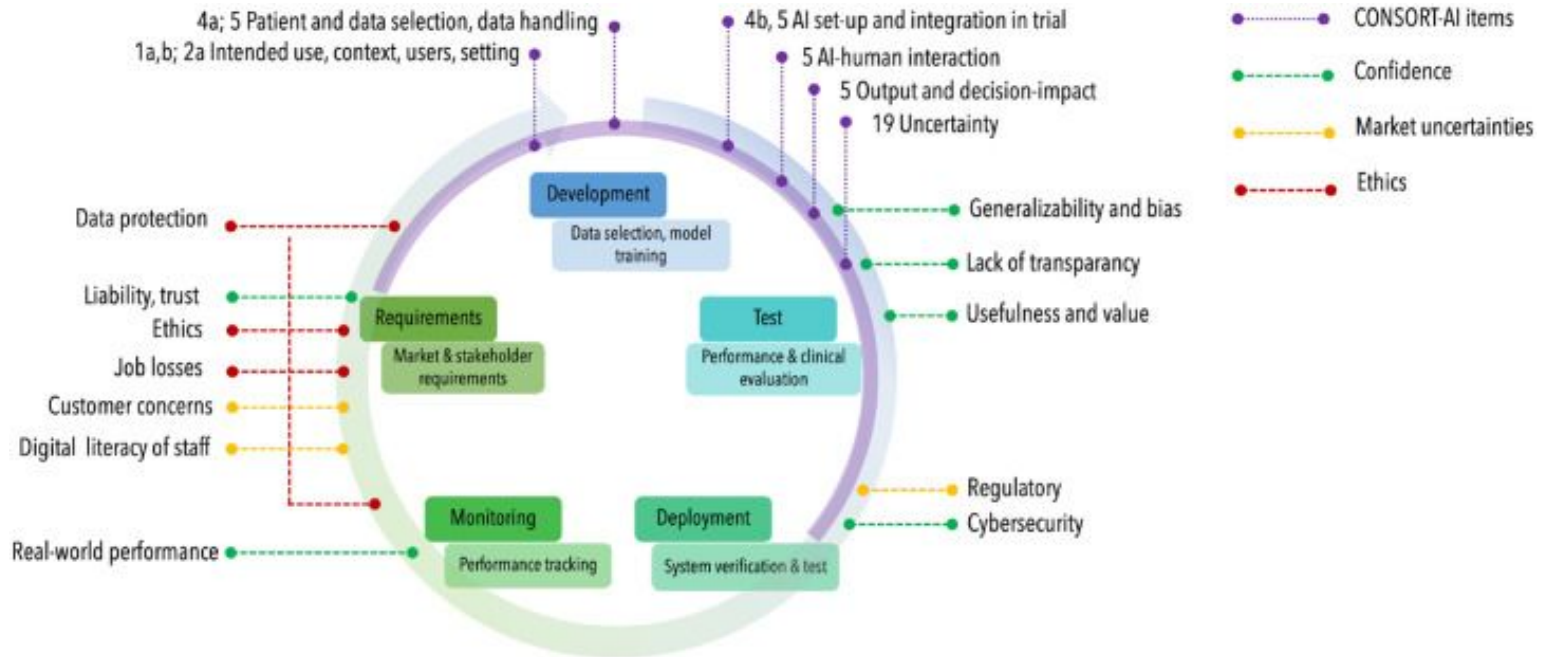
(tidak ada jawaban pasti)

Relevant virtues:

courage, compassion, justice

DEFICIENCY of VIRTUE (vice)	VIRTUE	EXCESS of VIRTUE (vice)
Cowardice	Courage	Rash
Insensible	Temperance	Dissipation
Stinginess	Generosity	Wastefulness
Chintzy	Magnificence	Vulgar
Aspersions	Magnanimity	Vainglory
Indolence	Industrious	Overambitious
Indifference	Caring	Controlling
Self-deprecation	Honest	Boastfulness
Boorishness	Charming	Buffoonery
Quarrelsome	Friendliness	Obsequious
Lying	Truthful	Tactless
Impatient	Tolerant	Doormat
Timid	Confident	Domineering
Fickle	Loyal	Gullible
Unsure	Vigilant	Impetuous
Cowardice	Protective	Bully
Fearful	Patient	Impulsive
Rudderless	Flexible	Rigid
Naïve	Practical	Cynical
Wimpy	Assertive	Arrogant
Selfish	Nurturing	Martyr
Paranoid	Confident	Arrogant
Pushover	Careful	Stubborn

Siklus hidup aplikasi AI



Penalaran Etis (*Ethical Reasoning*)

Definisikan masalah dan kumpulkan fakta

Aplikasikan prinsip-prinsip etika

Merancang aksi, evaluasi, dan refleksi

12 Langkah untuk Menganalisis Masalah Etika

1. Jelaskan isu etis yang ditemukan

2. Jabarkan fakta-fakta yang relevan

3. Identifikasi stakeholder

4. Perjelas nilai yang berlaku

5. Perhatikan konsekuensinya

6. Identifikasi kewajiban/tugas yang relevan

7. Refleksikan virtue yang berlaku

8. Perhatikan hubungan yang ada

9. Buat daftar respon yang potensial

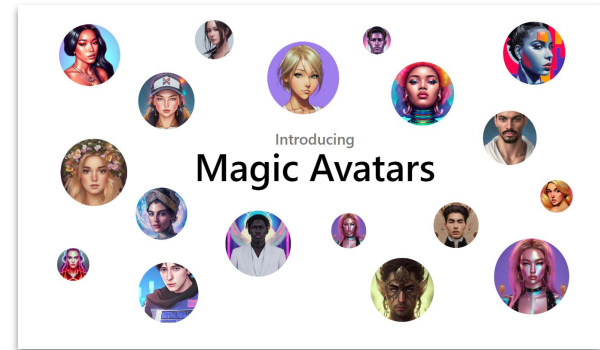
10. Gunakan imajinasi moral untuk menilai tiap respon

11. Pilih respon terbaik

12. Tinjau apa yang bisa diperbaiki

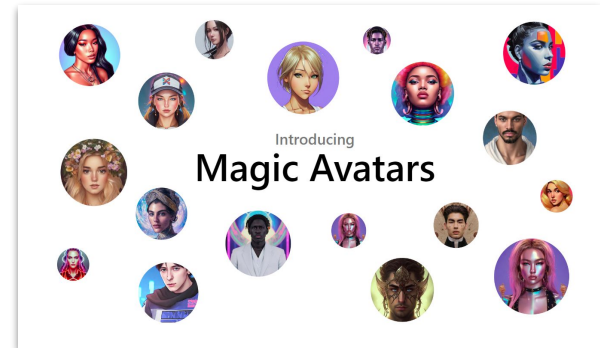
1. Jelaskan isu etis yang ditemukan

Lensa Prisma Labs, sebuah aplikasi pengeditan foto bertenaga kecerdasan artifisial, telah dituduh melakukan **pencurian karya seni** dan **eksploitasi seniman** sejak dirilisnya fitur Magic Avatar.



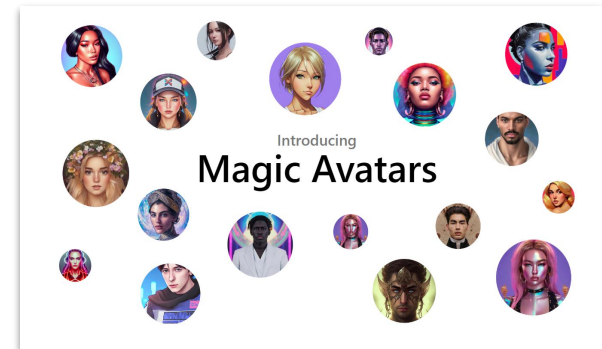
2. Jabarkan fakta-fakta yang relevan

- Lensa AI menjadi viral di awal tahun 2023 karena membuat orang penasaran dengan wajah buatan AI
- Generative AI yang digunakan adalah model Stable Diffusion 2 yang dilatih di LAION dataset dari gambar internet yang tidak terfilter
- Pelaku seni digital belum pernah mendapat kompensasi dari gambar yang “dicuri”
- Pelaku seni bisa “opt-out” dari database gambar yang digunakan untuk melatih model Stable Diffusion 3
- Belum ada kejelasan apakah Magic Avatar akan menggunakan model terbaru



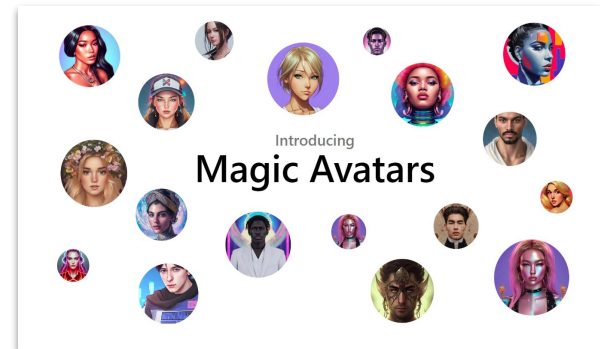
3. Identifikasi stakeholder

- Prisma Labs
 - Pengembang aplikasi Lensa AI
- Stability AI
 - Perilis model Stable Diffusion
- Komunitas seni digital
 - Yang merasa karya-karyanya “dicuri”
- Kompetitor
 - Yang bisa mengklaim telah memberikan keadilan untuk pelaku seni digital



4. Perjelas nilai yang berlaku

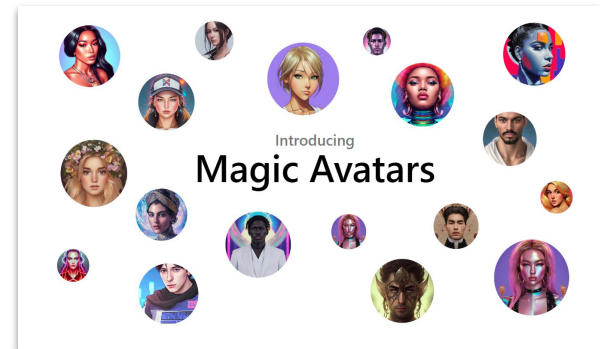
- Transparansi (transparency)
- Akuntabilitas (accountability)
- Keadilan (fairness)



5. Perhatikan konsekuensinya

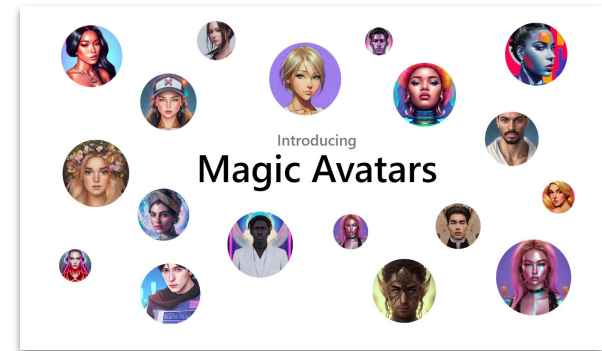
[dugaan pencurian karya seni dan eksploitasi seniman]

- Prisma Labs
 - *Public-image* yang buruk
 - Kehilangan **kepercayaan** pengguna
- Stability AI
 - *Backlash* dari komunitas seni digital
 - Penurunan reputasi sebagai pengembang AI
- Komunitas seni digital
 - Menuntut kompensasi
- Kompetitor
 - Bisa merancang strategi bisnis yang lebih baik



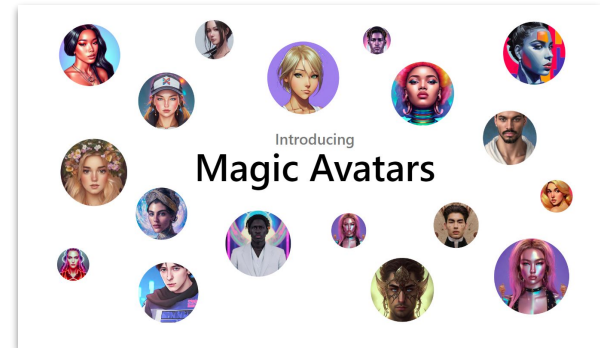
6. Identifikasi kewajiban/ tugas yang relevan

- Prisma Labs
 - Memberikan penjelasan mengenai bagaimana teknologi Magic Avatar dikembangkan
 - Menghargai hak kekayaan intelektual
- Stability AI
 - Memberikan penjelasan mengenai bagaimana model Stable Diffusion dikembangkan
 - Menghargai hak kekayaan intelektual
- Komunitas seni digital
 - waspada terhadap potensi pelanggaran hak cipta dan mengambil tindakan yang tepat jika karya mereka disalahgunakan
 - terlibat dalam diskusi konstruktif



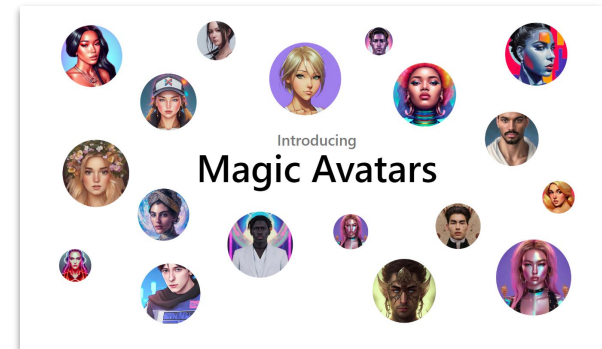
7. Refleksikan virtue yang berlaku

- Prisma Labs & Stability AI
 - Integritas
 - Transparansi
 - Tanggung jawab
 - Kolaborasi
- Komunitas seni digital
 - Integritas
 - Keberanian
 - Advokasi
 - Kolaborasi



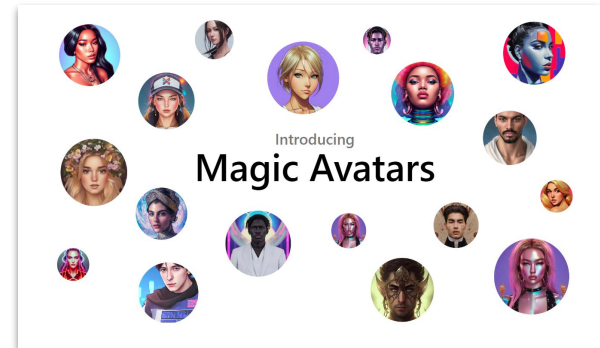
8. Perhatikan hubungan yang ada

- Prisma Labs & Stability AI
 - AI as a service
 - Memegang prinsip etika yang sama
- Stability AI & Komunitas seni digital
 - (harusnya) Simbiosis mutualisme



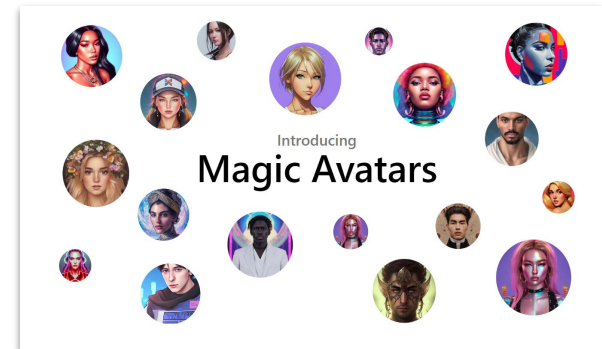
9. Buat daftar respon yang potensial

- Prisma Labs
 - a. Mengeluarkan pernyataan
 - b. Mengubah praktik bisnis/desain teknis
- Stability AI
 - a. Mengeluarkan pernyataan
 - b. Mengkompensasi pelaku seni digital
- Komunitas seni digital
 - a. Melakukan advokasi legislatif
 - b. Menuntut Stability AI secara hukum



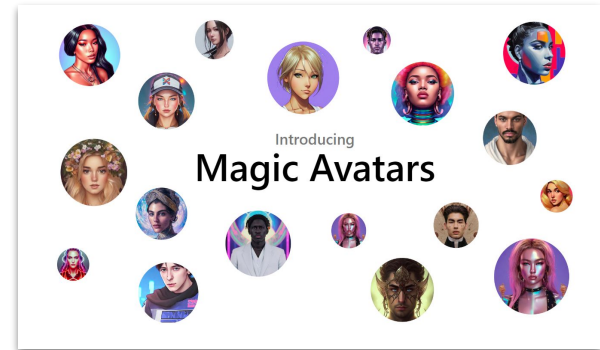
10. Gunakan imajinasi moral untuk menilai tiap respon

- Prisma Labs
 - a. Mengeluarkan pernyataan?
 - b. Mengubah praktik bisnis/desain teknis?
- Stability AI
 - a. Mengeluarkan pernyataan?
 - b. Mengkompensasi pelaku seni digital?
- Komunitas seni digital
 - a. Melakukan advokasi legislatif?
 - b. Menuntut Stability AI secara hukum?



11. Pilih respon terbaik

- Prisma Labs
 - a. Mengeluarkan pernyataan?
- Stability AI
 - a. Mengeluarkan pernyataan?
 - b. Mengkompensasi pelaku seni digital?
- Komunitas seni digital
 - a. Melakukan advokasi legislatif?



12. Tinjau apa yang bisa diperbaiki untuk kedepannya

- Prisma Labs
 - a. Melakukan kajian komprehensif bagaimana produk yang mereka keluarkan akan memengaruhi pihak-pihak yang relevan
- Stability AI
 - a. Memiliki sistem terpercaya dan berkelanjutan di mana pelaku seni bisa memilih agar karya-karyanya tidak dilibatkan dalam pelatihan model generative AI
 - b. Memengaruhi praktik industri generative AI
- Komunitas seni digital
 - a. Mengadvokasi regulasi generative AI

Dream Bigger with Adobe Firefly.

Experiment, imagine and create an infinite range of images with Firefly, generative AI-powered content creation from Adobe.

OUR APPROACH TO GENERATIVE AI

Creators first.

Adobe is committed to developing creative generative AI responsibly, with creators at the centre. Our mission is to give creators every advantage — not just creatively, but practically. As Firefly evolves, we will continue to work closely with the creative community to build technology that supports and improves the creative process.

Pelajaran untuk “kompetitor” - Adobe Firefly

Trustworthy AI

Sistem AI yang dapat dipercaya artinya sistem tersebut **dapat diandalkan** untuk beroperasi dengan cara yang selaras dengan nilai-nilai kemanusiaan, prinsip etika, dan norma masyarakat.

Trustworthy AI

(GDPR)

Pengembangan dan pengaplikasian AI yang

- **lawful:** patuh terhadap semua hukum dan peraturan yang berlaku
- **ethical:** menghormati prinsip dan nilai etika
- **robust:** dapat diandalkan dari segi teknis dan tetap memperhatikan sosial-lingkungan

7 Persyaratan Utama Sistem AI Yang Dapat Dipercaya (Trustworthy)

(GDPR)

1. Agensi dan pengawasan manusia
2. *Robustness* dan keamanan teknis
3. Privasi dan tata kelola data
4. Transparansi
5. Keanekaragaman, non-diskriminasi, dan keadilan
6. Kesejahteraan masyarakat dan lingkungan
7. Akuntabilitas

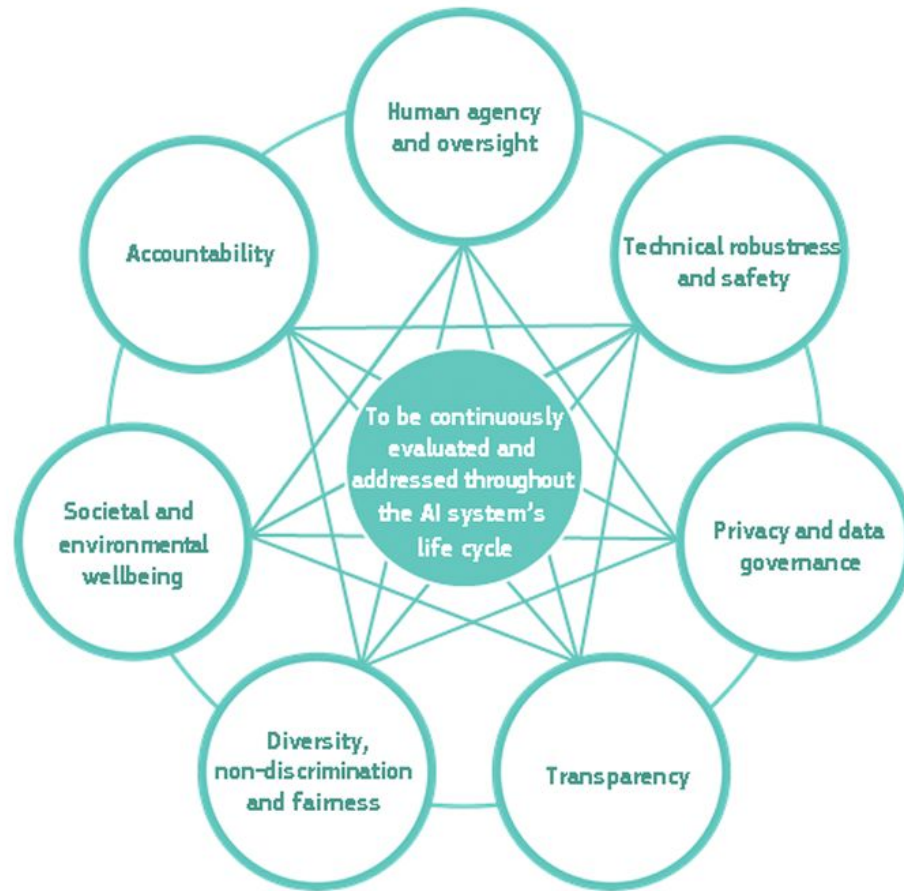


Figure 2: Interrelationship of the seven requirements: all are of equal importance, support each other, and should be implemented and evaluated throughout the AI system's lifecycle

Read Along (Bolo)

Voice and AI design to grow child literacy



7 Persyaratan Utama Sistem AI: Studi Kasus Bolo

1. Agensi dan pengawasan manusia
 - a. Menyadari bahwa orangtua adalah pengawas anak → *onboarding* untuk orangtua
 - b. Meminta izin untuk merekam suara
2. *Robustness* dan keamanan teknis
 - a. Menyadari bahwa performa *speech model* belum sempurna → konsiderasi UX app
 - b. *Precision and recall tradeoff*
3. Privasi dan tata kelola data?
4. Transparansi?
5. Keanekaragaman, non-diskriminasi, dan keadilan?
6. Kesejahteraan masyarakat dan lingkungan?
7. Akuntabilitas?

Key Takeaways

- Semakin tingginya kemampuan sistem AI, semakin banyak isu tantangan kerugian yang signifikan secara etis
- Perkembangan ilmu etika teknologi dan etika AI terkadang belum bisa mengikuti kemajuan sisi teknis
 - Apalagi perkembangan hukum dan legal
- Sebagai *engineer*, mulailah bangun kesadaran etika dengan mengikuti perkembangan terkini dan melakukan *case study*
- Menerjemahkan *framework* menjadi aksi adalah suatu tantangan tersendiri

Resources

EU GDPR - Trustworthy AI Guidelines:

<https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>

KPMG & University of Queensland - Model of Trustworthy AI:

<https://kpmg.com/au/en/home/insights/2020/11/trustworthy-ai.html>

Google - People + AI Research: <https://pair.withgoogle.com/>

Terima kasih!