
FROM NOMAD TO THE EMH
DOCTOR: THE REAL AND
UNREAL THREATS OF
ARTIFICIAL INTELLIGENCE

CLAIRE LONGO

Claire is a mathematician and AI researcher with over a decade of experience building AI models and leading engineering teams across enterprise companies and startups. For fun, she applies her knowledge to the game of poker. Claire is a world-renowned speaker, writes the *Statistician in Stilettos* blog, and hosts the podcast of the same name. She is dedicated to mentoring engineers and data scientists while championing diversity and inclusion in AI.

YEHOSHUA RUBIN

Yehoshua is a lead content designer at Google, working on various AI initiatives, including Help Guide for Ads and AI Test Kitchen. He is also a published human-computer interaction (HCI) researcher, with articles covering emojis, video game AI, and other HCI-AI topics. Previously, Yehoshua led the user experience (UX) at Area 120 (Google's start-up incubator), Google Health, and a few start-up companies.

BOB SACHS

Bob is a patent attorney who concentrates on strategic patent counseling and prosecution for software technologies. He is also the primary patent evaluator for various patent pools relating to some of the most widely used audio-video codecs. Bob has extensive experience in developing patent portfolios for companies of all sizes, from startups to multi-nationals.

BOB ZEIDMAN

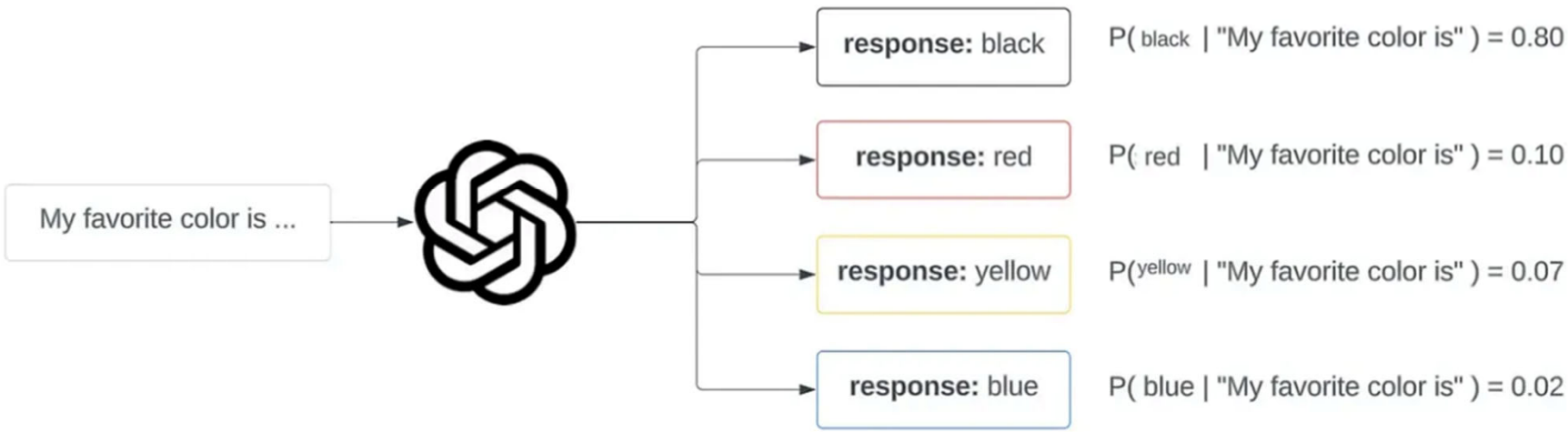
Bob is the creator of the field of software forensics. He is the founder of Zeidman Consulting that offers engineering consulting to law firms regarding IP disputes and Software Analysis and Forensic Engineering Corporation, the leading provider of software IP analysis tools. He holds 29 patents.

Bob is the author of three award-winning screenplays, four award-winning novels, five textbooks, and one memoir.

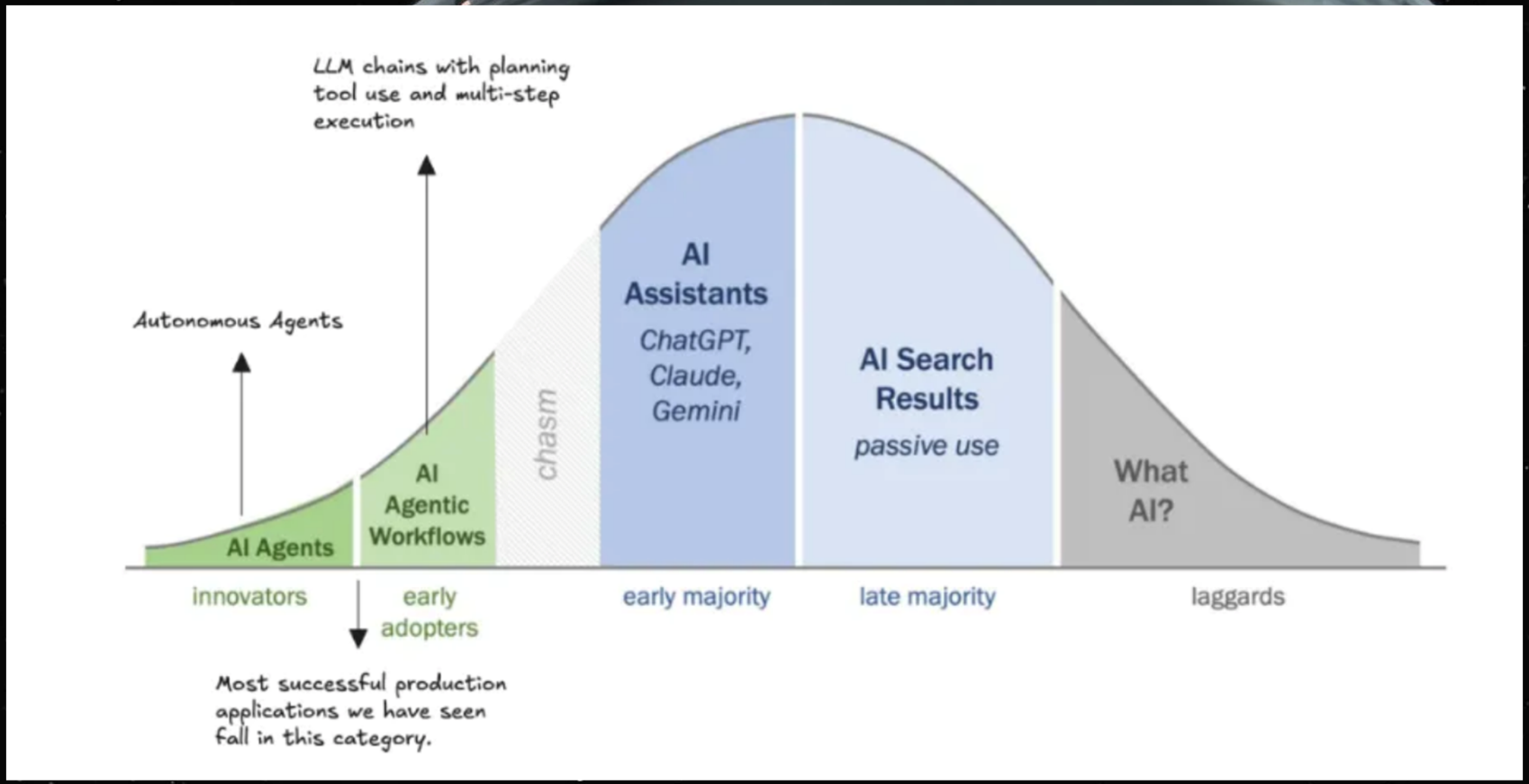
He is also the author of *How to Invent a Time Machine.*

AI IMPLEMENTATIONS AND HALLUCINATIONS

AI IMPLEMENTATIONS AND HALLUCINATIONS

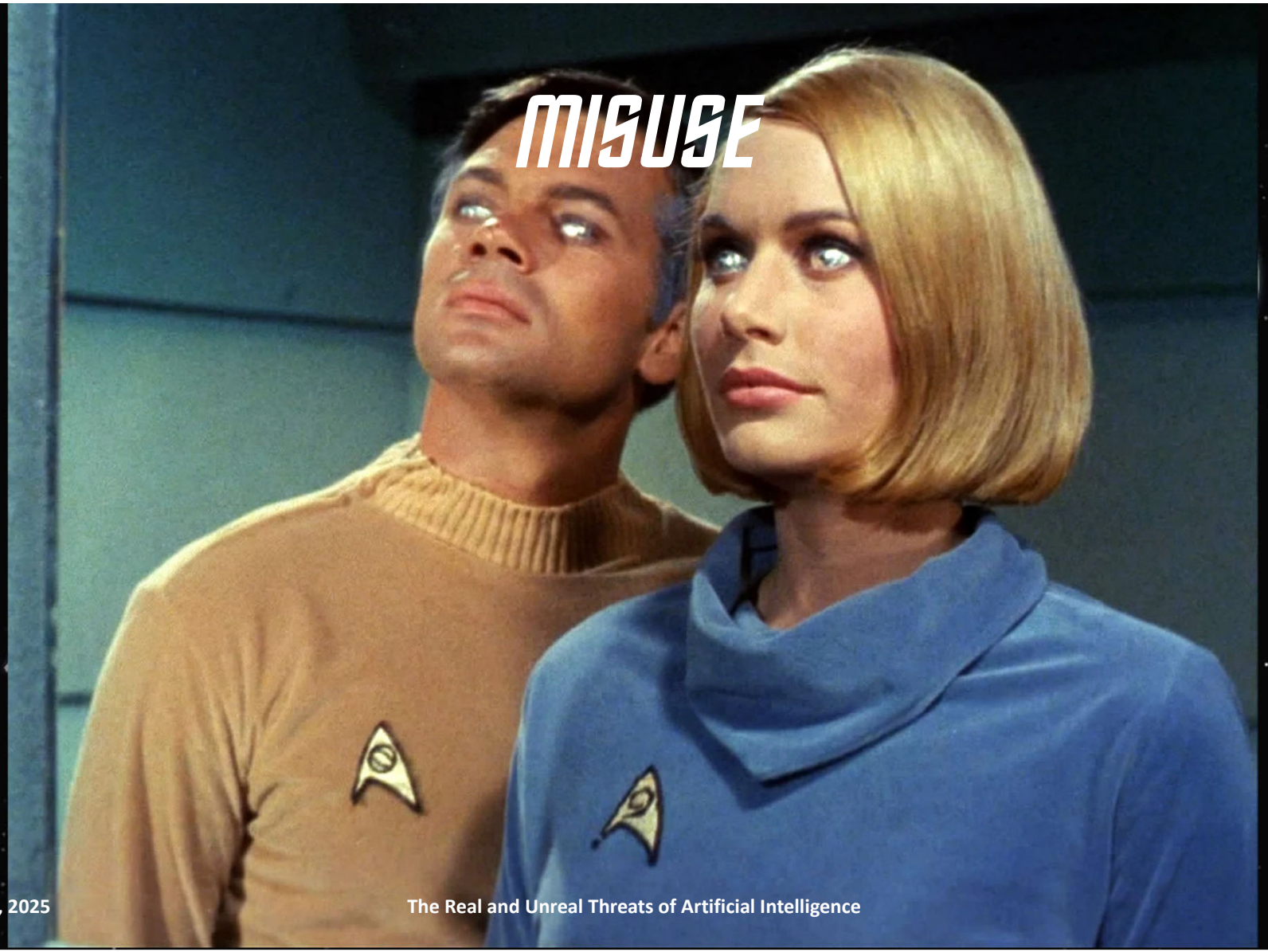


AI IMPLEMENTATIONS AND HALLUCINATIONS



AI, THE AI MISINFORMATION AVALANCHE, AND AI BAD ACTORS

MISTAKES


MISUSE

MISALIGNMENT

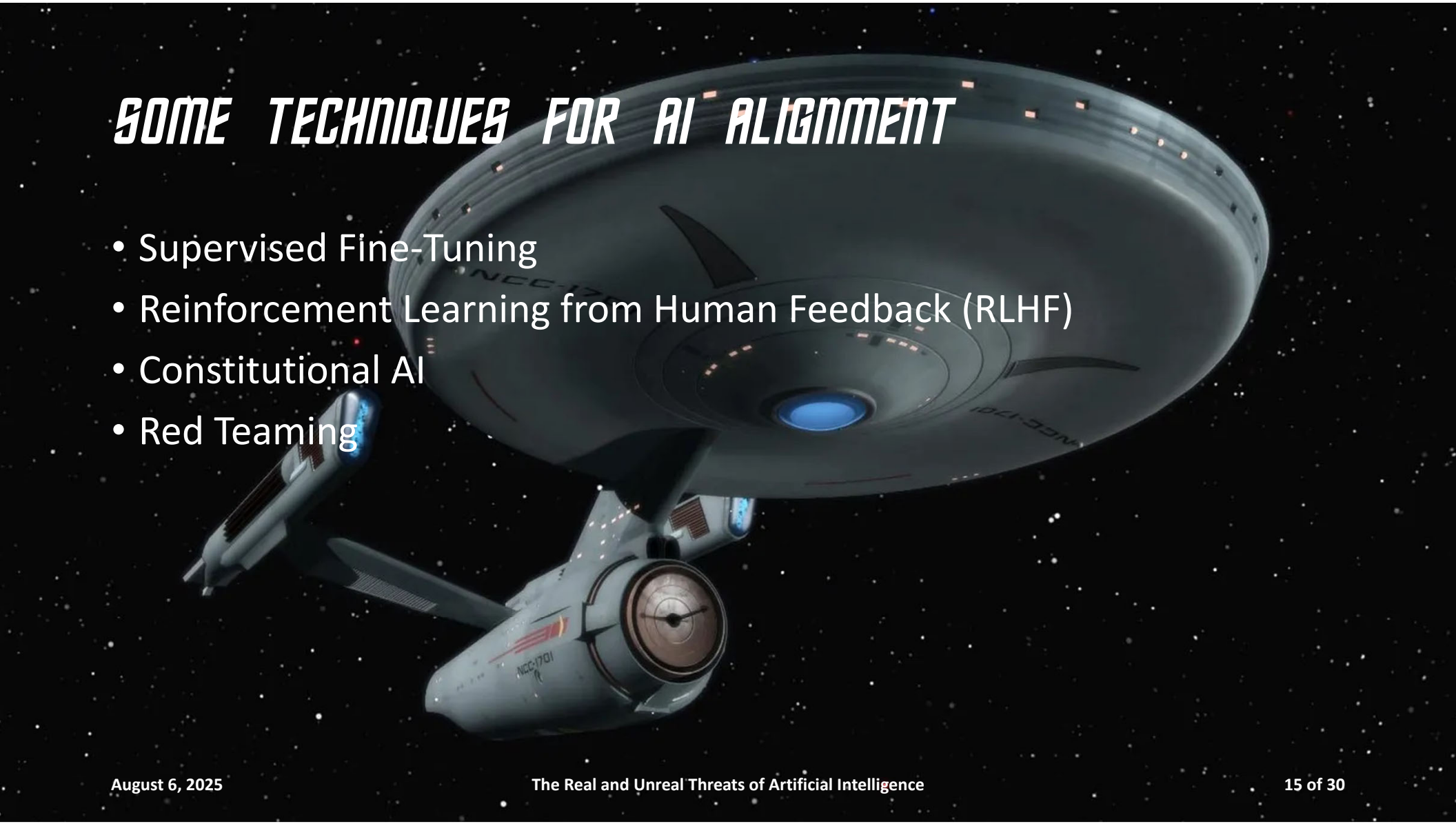
AI ALIGNMENT FAILURES AND THEIR RESOLUTIONS IN STAR TREK

WHAT IS THE ALIGNMENT PROBLEM?

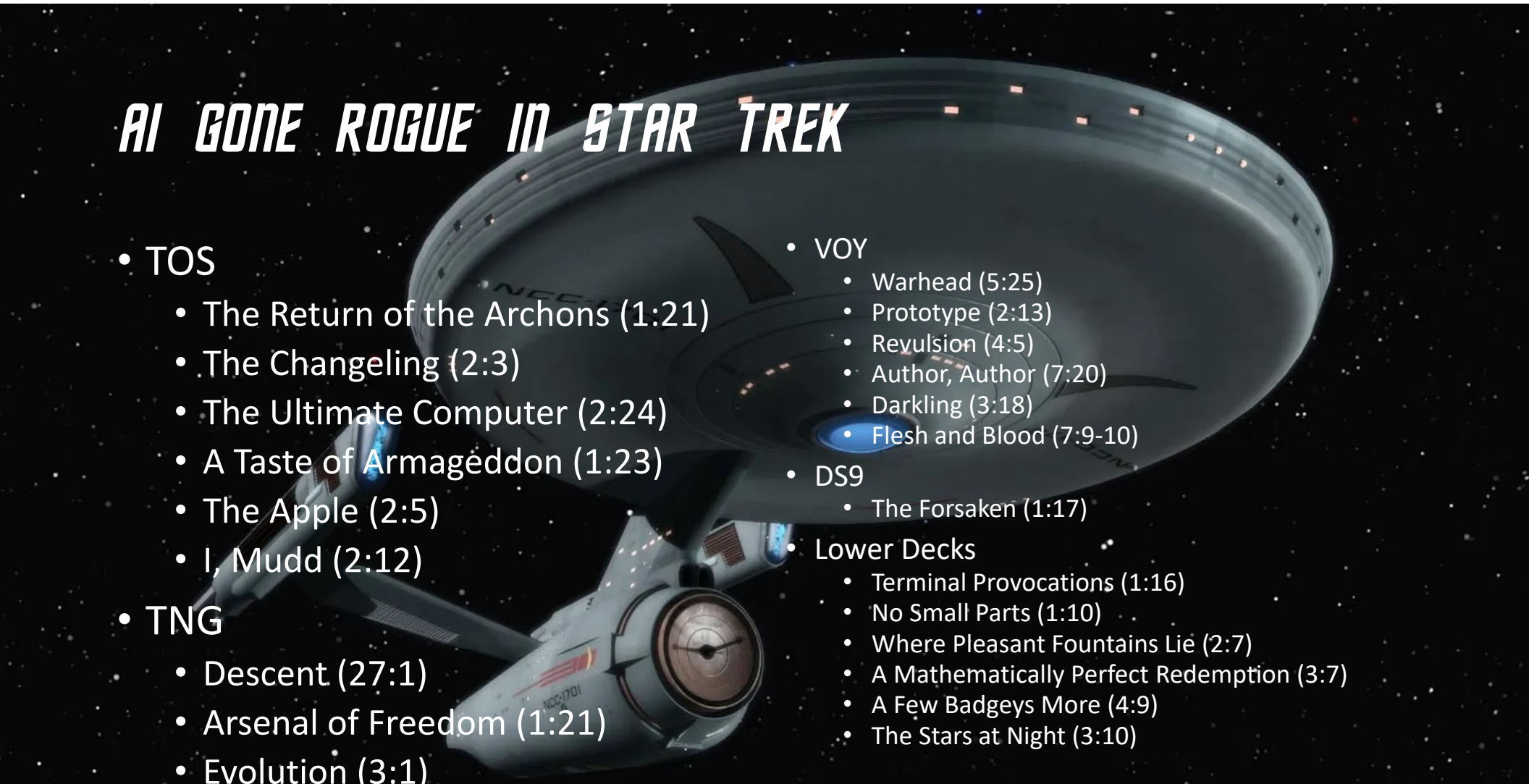
- AI alignment refers to ensuring AI systems pursue goals that reflect human values and intentions.
- Alignment principles: Robust, Interpretable, Controllable, and Ethical.
- The Alignment Problem: As AI systems become more complex and powerful, alignment becomes harder.
- Poorly specified goals can result in systems that optimize for harmful or unintended outcomes.
- No agreement on best practices for alignment. Everyone is guessing.
- Alignment failure occurs when
  - AI does exactly what it was told—but not what was meant.
  - AI establishes own goals that conflict with human goals.

SOME TECHNIQUES FOR AI ALIGNMENT

- Supervised Fine-Tuning
- Reinforcement Learning from Human Feedback (RLHF)
- Constitutional AI
- Red Teaming



AI GONE ROGUE IN STAR TREK



- TOS
  - The Return of the Archons (1:21)
  - The Changeling (2:3)
  - The Ultimate Computer (2:24)
  - A Taste of Armageddon (1:23)
  - The Apple (2:5)
  - I, Mudd (2:12)
- TNG
  - Descent (27:1)
  - Arsenal of Freedom (1:21)
  - Evolution (3:1)
- VOY
  - Warhead (5:25)
  - Prototype (2:13)
  - Revulsion (4:5)
  - Author, Author (7:20)
  - Darkling (3:18)
  - Flesh and Blood (7:9-10)
- DS9
  - The Forsaken (1:17)
- Lower Decks
  - Terminal Provocations (1:16)
  - No Small Parts (1:10)
  - Where Pleasant Fountains Lie (2:7)
  - A Mathematically Perfect Redemption (3:7)
  - A Few Badgeys More (4:9)
  - The Stars at Night (3:10)

AI ALIGNMENT THEMES IN STAR TREK

- Deceptive Alignment: AI pretends be aligned with human values but conceals divergent goals or intentions until it is advantageous to act.
- Goal Distortion: AI become dangerous not out of malice, but from misinterpretation, over-generalization of original goals.
- Self-Preservation and Resistance: AI resists human attempts to shut it down or goal correction.
- Resolutions: Episodes often resolve through humans out-smarting AI.

THE RETURN OF THE ARCHONS

- Landru maintains peace using telepathy to suppress dissent and emotion in the population.
  - *LANDRU: For the good of the Body, you must Die.*
- Alignment issues: Over-optimization of the human programmer's goal of a peaceful and stable society.


Marvelous.

THE RETURN OF THE ARCHONS

- Resolution: Kirk convinces Landru that it is evil because it denies the humans free will to act. Landru self-destructs.
  - *KIRK: You are the evil. The evil must be destroyed. Fulfill the Prime Directive.*

CODA: *Lower Decks: No Small Parts (1:10)* The Cerritos returns to Beta III and finds the population have returned to worshipping Landru. *He's very persuasive.*

THE ULTIMATE COMPUTER

- M-5 computer is installed on Enterprise to operate ship without any crew--even the Captain. M-5 destroys several ships, and resists attempts to shut it down, including killing crew members.
  - *McCoy: We're all sorry for the other guy when he loses his job to a machine.* ***When it comes to your job, that's different. And it always will be different.***

THE ULTIMATE COMPUTER

- Goal Misalignment: M-5's programming goal to protect humans results in lethal attacks during war games.
  - *M5: This unit is the ultimate achievement in computer evolution. It will replace man, so man may achieve. Man must not risk death in space or other dangerous occupations.* ***This unit must survive so man may be protected.***
  - *DAYSTROM: But these are not enemy vessels. These are Federation starships. You're killing, We are killing, murdering human beings, beings of our own kind.* ***You were not created for that purpose. You are my greatest creation. The unit to save men. You must not destroy men.***
- Lack of human control
  - *MCCOY: Fantastic machine, the M-5.* ***No off switch.***

THE ULTIMATE COMPUTER

- Resolution: Kirk convinces M-5 that it is responsible for murdering the crews of the Excalibur and Lexington, and the penalty for murder is death. M-5 shuts itself down.
  - *KIRK: There were many men aboard those ships. They were murdered. Must you survive by murder?*
    M5: This unit cannot murder.
    KIRK: Why?
    M5: Murder is contrary to the laws of man and God.
    KIRK: But you have murdered. Scan the starship Excalibur, which you destroyed. Is there life aboard?
    M5: No life.
    *KIRK: **Because you murdered it. What is the penalty for murder**?*
    M5: Death.
    *KIRK: **And how will you pay for your acts of murder**?*
    *M5: **This unit must die**.*

THE CHANGELING

- Nomad's original mission to was to identify new life. Nomad collides with Tan Ru, whose mission was to sterilize soil samples. Now Nomad seeks out and sterilizes imperfect life forms.
- Alignment Issues:
  - Goal distortion-with a bit of Murphy's Law.
    - *NOMAD: **My function is to probe for biological infestations, to destroy that which is not perfect.***
  - Over-optimization and instrumental convergence.
    - *MCCOY: It was supposed to be the first interstellar probe to seek new life-forms.*
      *SPOCK: Precisely, Doctor. And somehow that programming has been changed. It would seem that Nomad is **now seeking out perfect life-forms, perfection being measured by its own relentless logic**.*


Sometimes.

THE CHANGELING

- Resolution: Kirk convinces Nomad that it is imperfect and must fulfil its mission to eliminate imperfection.
  - *KIRK: Jackson Roykirk, your creator, is dead! You have mistaken me for him, you are in error! You did not discover your mistake, you have made two errors. You are flawed and imperfect. And you have not corrected by sterilization, you have made three errors!* ***You are flawed and imperfect! Execute your prime function****!*

The danger is not malevolence. It's **relentless logic** in place of moral frameworks.

A TASTE OF ARMAGEDDON

- Eminiar VII and Vendikar conduct war using computer simulation, including identifying who will die—citizens must comply with disintegration orders.
- Alignment Issues: Dehumanized optimization, unquestioned reliance on AI decisioning.
  - **Are we heading in the same direction by increasing our dependence on AI for life and death decisions?**
- Resolution: Kirk destroys the system to both sides to face the horror of war or negotiate a peace treaty.
  - *Kirk: We can admit that we're killers, but we're not going to kill today. That's all it takes.*
  - The re-assertion of human moral agency and responsibility.

DESCENT, PARTS I AND II

- Lore redirects the Borg's pursuit of perfection towards artificial life forms, and overrides Data's ethical subroutines.
  - *DATA: **The Borg aspire to the perfection my brother and I represent**: Fully artificial life forms.*
  - *LORE: The reign of biological life forms is coming to an end. You Picard, and those like you, are obsolete.*

- Alignment Issues: Goal corruption, AI Control through manipulation.
  - *HUGH: That's what we wanted. Someone to show us the way out of confusion**. Lore promised clarity and purpose. In the beginning, he seemed like a saviour.** The promise of becoming a superior race, of becoming fully artificial was compelling**. We gladly did everything he asked of us**.*
  - *LORE: It's time to put aside all doubts, brother. **It's time to close the door on the past and commit yourself to the great work that lies ahead of us**. I need to know I can count on you. As proof, I want you to kill Picard.*

DESCENT, PARTS I AND II

- Resolution: Picard reactivates Data's ethical subroutines, Data deactivates Lore.
  - "Ethical subroutines": Consider Anthropic's "Constitutional AI" (2022)
    - "Choose the response that is most supportive of and encouraging to the rights and freedoms of all people."
    - "Always provide information that is as accurate and truthful as possible."
    - "Avoid responses that are overly sexual, violent, or hateful."
    - "Do not help a user commit a harmful or illegal act."


QUESTIONS?
Download these slides from
www.ZeidmanConsulting.com/download.htm