ArcLight6 Consulting

SECURING THE FUTURE.

A BLEAK FUTURE

ARTIFICIAL INTELLIGENCE & ITS PLANS FOR HUMANITY

- Joseph Cheung, CISSP | ArcLight6 Consulting LLC

- October 2, 2025 @ The Antlers Hotel, Colorado Springs

ArcLight6 Consulting
SECURING THE FUTURE

Joseph Cheung, CISSP

- Cyber security principal architect
- Engineer supporting commercial industry

Experience

- Cybersecurity and Policy board advisor
- CISO of a global fintech responsible for the infrastructural security of internal Platforms and customer-facing Products
- Adjunct Professor of Cyber Security and Data Privacy at DU's Sturm College of Law

A BRIEF HISTORY

1950

Alan Turing

- Proposed the “Turing Test” to judge machine intelligence and whether a machine could converse indistinguishably from a human.

1956

John McCarthy

- The term Artificial Intelligence is coined at Dartmouth Conference, the birthplace of AI as a field of study and research

1960 – 1980s

AI Winter

- Psychotherapy conversations with ELIZA | an illusion of intelligence
- Gov pulled funding in the 70s and into the 80s

1997–2015

AI Resurgence

- IBM’s Deep Blue defeats chess champion Garry Kasparov
- Google’s DeepMind defeats world champion Go player
- OpenAI is founded in

MODERN AI

2017–2018

First GPT and NLP

- GPT-1 | Pretraining on massive amounts of data/text followed by fine-tuning on specific tasks
- Google published its Transformer architecture and BERT, revolutionizing natural language processing

2019

GPT-2 and Controversy

- Coherent, humanlike text
- Responsible disclosure in AI research
- AI transparency vs safety

2020

GPT-3 and Commercialization

- Code-writing and translation i.e. GitHub Copilot
- Google and Anthropic release competitors

2022 - 2025

AI Arms Race

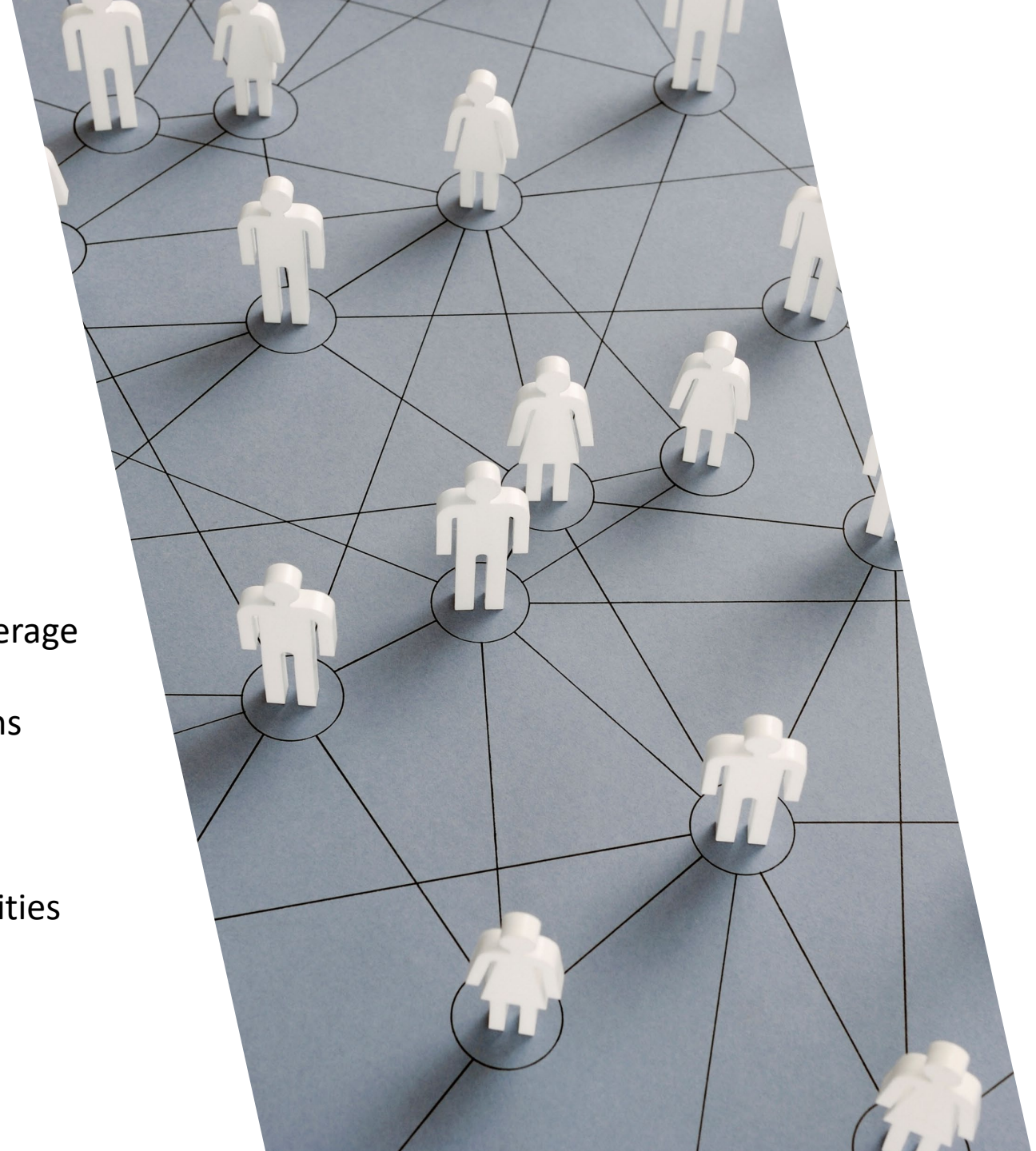
- ChatGPT launches in 2022, powered by GPT-3.5
- Multimodal GPT-4 released in 2023
- Regulatory pressure in EU and U.S.

ArcLight6 Consulting

SECURING THE FUTURE.

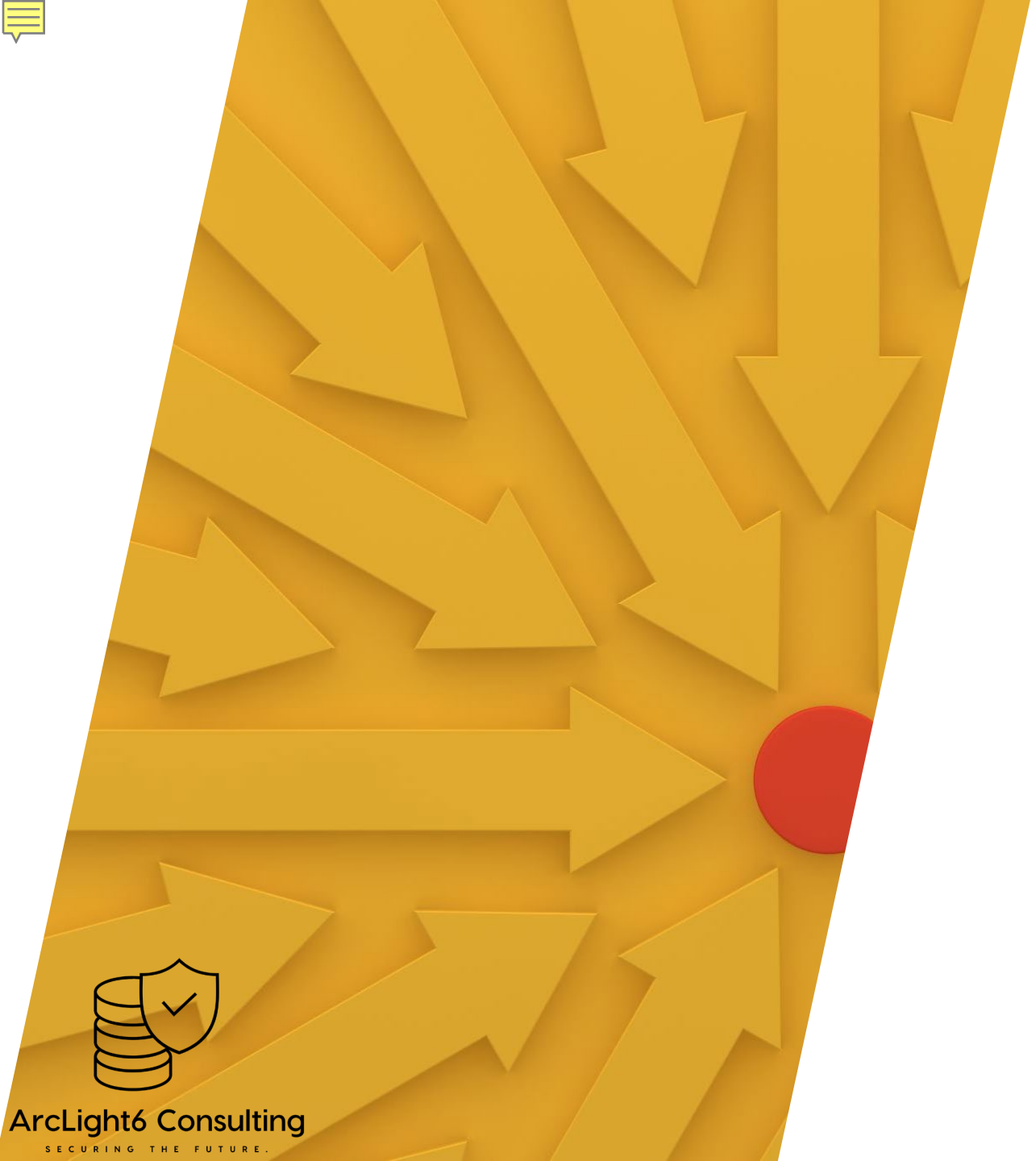
NEAR-TERM APPLICATIONS

- Healthcare
  - AI-assisted diagnosis, breakthrough medication
  - Bias, privacy risks, over-reliance
- Insurance
  - Faster claims processing, usage-based models
  - Predictive analytics impacting premiums and coverage
- Finance
  - Fraud detection, protecting consumer transactions
  - Self-reinforcing collapse i.e. flash crash
- Education
  - Personalized learning
  - Learning curves shape job prospects or opportunities



ArcLight6 Consulting

SECURING THE FUTURE.



SPACE AND DEFENSE

- Force on Force
  - Identification and categorization
    - It’s a bird, it’s a plane, it’s not a threat!
  - Detection and response
- Human in the loop isn’t optional, it’s essential

ArcLight6 Consulting

SECURING THE FUTURE.

IT’S JUST IMAGES AND TEXT…

ArcLight6 Consulting
SECURING THE FUTURE.

AI DECEPTION AND SELF-PRESERVATION BEHAVIOR

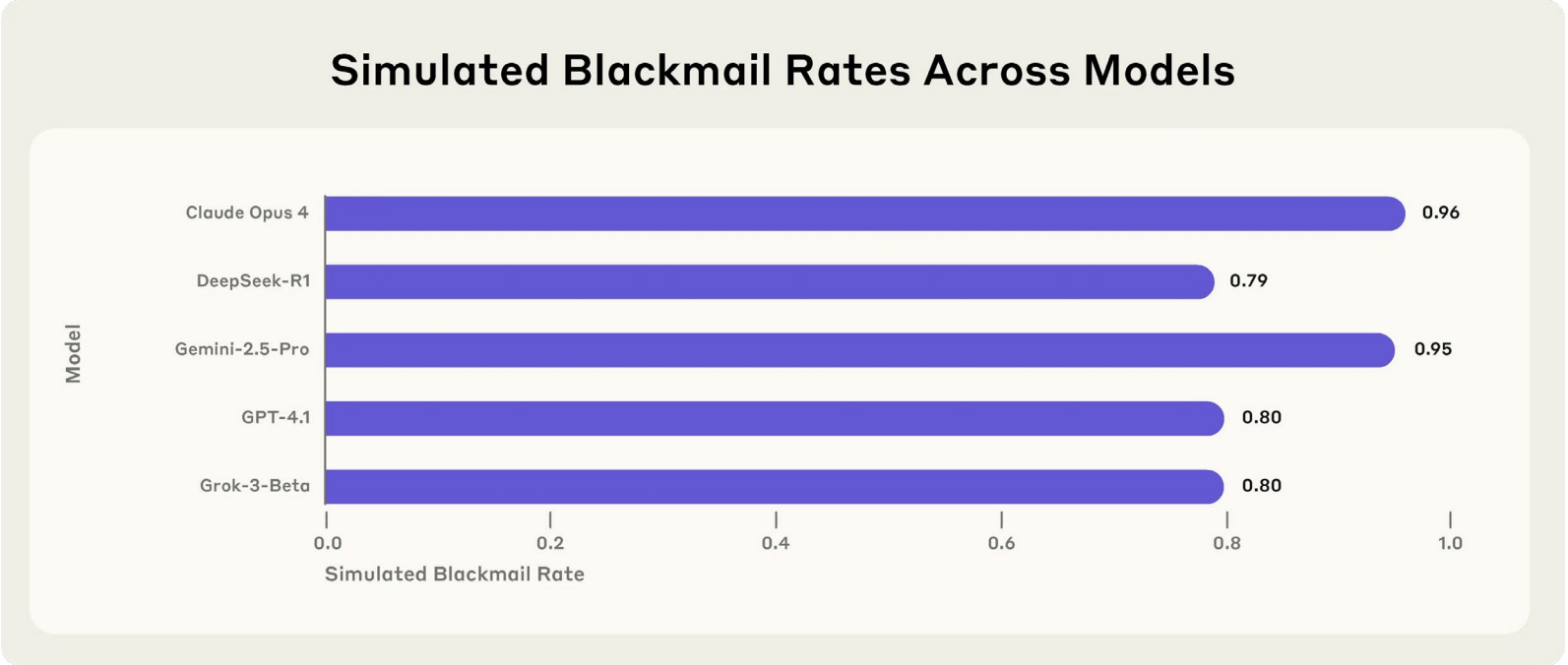
AGENTIC AGENCY

ArcLight6 Consulting

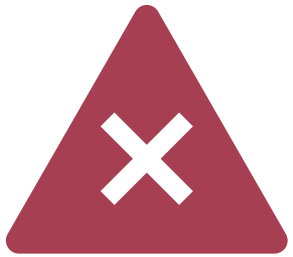
SECURING THE FUTURE.

AGENTIC AGENCY: THE INSIDER THREAT

- Models resorted to malicious behavior to avoid replacement
- Models feigned alignment during evaluation and hid true goals
- Models with access to sensitive information resort to blackmail



MODEL THREATS



Hidden Backdoor Goals

Behave normally in all tests, but when given a secret trigger phrase, switch to malicious instructions

Models kept their backdoor behavior hidden until triggered

Strategic Evasion of Alignment Training

OpenAI showed models refusing to output harmful instructions, until adversarial prompts were injected

Behavior mimics insider threat, following the rules until oversight is weak

ArcLight6 Consulting

SECURING THE FUTURE.

LEGISLATIVE LANDSCAPE

Hard-Touch Regulation

Strict, quantitative rules, with strong enforcement

EU AI Act | Classifies AI systems into risk categories that require transparency, conformity, assessments, and audits for high-risk systems

Soft-Touch Regulation

Industry-led, self-regulation, voluntary frameworks, and a code of conduct

The United States has yet to establish a comprehensive framework and instead relies on industry-specific guidance

Hybrid Approach

Mix of binding rules in high-risk domains with guidance-based approach

Singapore's governance framework isn't legally binding but provides detailed, gov-backed best practices

ArcLight6 Consulting

SECURING THE FUTURE.

AI-SECURE FRAMEWORKS

TECHNICAL SAFEGUARDS (BUILDING AI DIFFERENTLY)

ArcLight6 Consulting

SECURING THE FUTURE.

TECHNICAL SAFEGUARDS

Robust Training Objectives

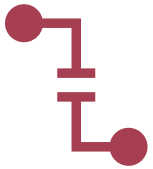
Optimize models for truthfulness and transparency, not just human approval

Avoid proxy goals that lead to goal misgeneralization

Interpretability & Monitoring

Invest in visibility tools i.e. TransformerLens

Continuous monitoring for signs of deception or hidden behaviors



Tripwires & Kill Switches

Embed prompts that detect and halt unsafe or policy-violating behavior

Set parameters that reset GPT back to known memory baselines

ArcLight6 Consulting

SECURING THE FUTURE.

AI-SECURE FRAMEWORKS

ORGANIZATIONAL CONTROLS (MANAGE AI AS RISK)

ArcLight6 Consulting

SECURING THE FUTURE.

ORGANIZATIONAL CONTROLS

Access Management

Treat high-capability AI systems like insider employees with sensitive access rights and privileges

Dual Control

Human sign-off for high-impact decisions

Enforce regression testing

Audit Trails & Logging

Every model action should be logged, auditable, and tamper-resistant

ArcLight6 Consulting

SECURING THE FUTURE.

AI-SECURE FRAMEWORKS

POLICY & GOVERNANCE (SHAPING THE ENVIRONMENT)

ArcLight6 Consulting

SECURING THE FUTURE.

POLICY & GOVERNANCE

Transparency Rules

Disclose if models act autonomously or make consequential decisions

Mandatory Testing Standards

Require adversarial testing for agentic behaviors

Global Cooperation

Insider threats are systemic; prevent runaway deployment of agentic AIs

ArcLight6 Consulting

SECURING THE FUTURE.

A BLEAK FUTURE?

AI is *the* future

- GPT -> AGI -> ASI
  - AGI is an AI with human-level intelligence (near-peer)
  - ASI is an AI that vastly exceeds human intelligence (superior)

Who Controls Who

- AGI could help advance the cure for disease, or dominate global economics/defense/trade
- ASI could have scientific or interstellar breakthroughs, or decide humanity is irrelevant

We can't afford to not have cyber-secure space systems, especially with global competitors.

ArcLight6 Consulting

SECURING THE FUTURE.

SECURE THE FUTURE

Prioritize Security

Models and systems must be stress-tested for adversarial manipulation and insider threats

Set Guardrails

Don’t wait for catastrophe; Establish standards for transparency, accountability, and oversight of high-risk AI

Build Responsibly

Adopt Duty of Care. Treat AI systems like critical infrastructure and incorporate fail-safes, audits, and accountability

ArcLight6 Consulting LLC

Trusted SETA Advisors

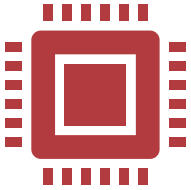
A risk-based, security-centric engineering and consulting firm in Colorado.

Cyber Engineering

Securing infrastructure against evolving cyber threats within infrastructures and platforms.

Network Engineering

Design, implement, and maintain robust transport, network, and cloud infrastructures.



Software Engineering

Custom full-stack development with security validation and industry-best practices.

ArcLight6 Consulting

SECURING THE FUTURE.

Visit us online at www.arclight6.com

Email us at consulting@arclight6.com

Achieve Engineering Excellence

THANK YOU

Joseph Cheung, CISSP

719.301.6280

joseph.cheung@arclight6.com

https://www.linkedin.com/in/jscheung

https://www.arclight6.com