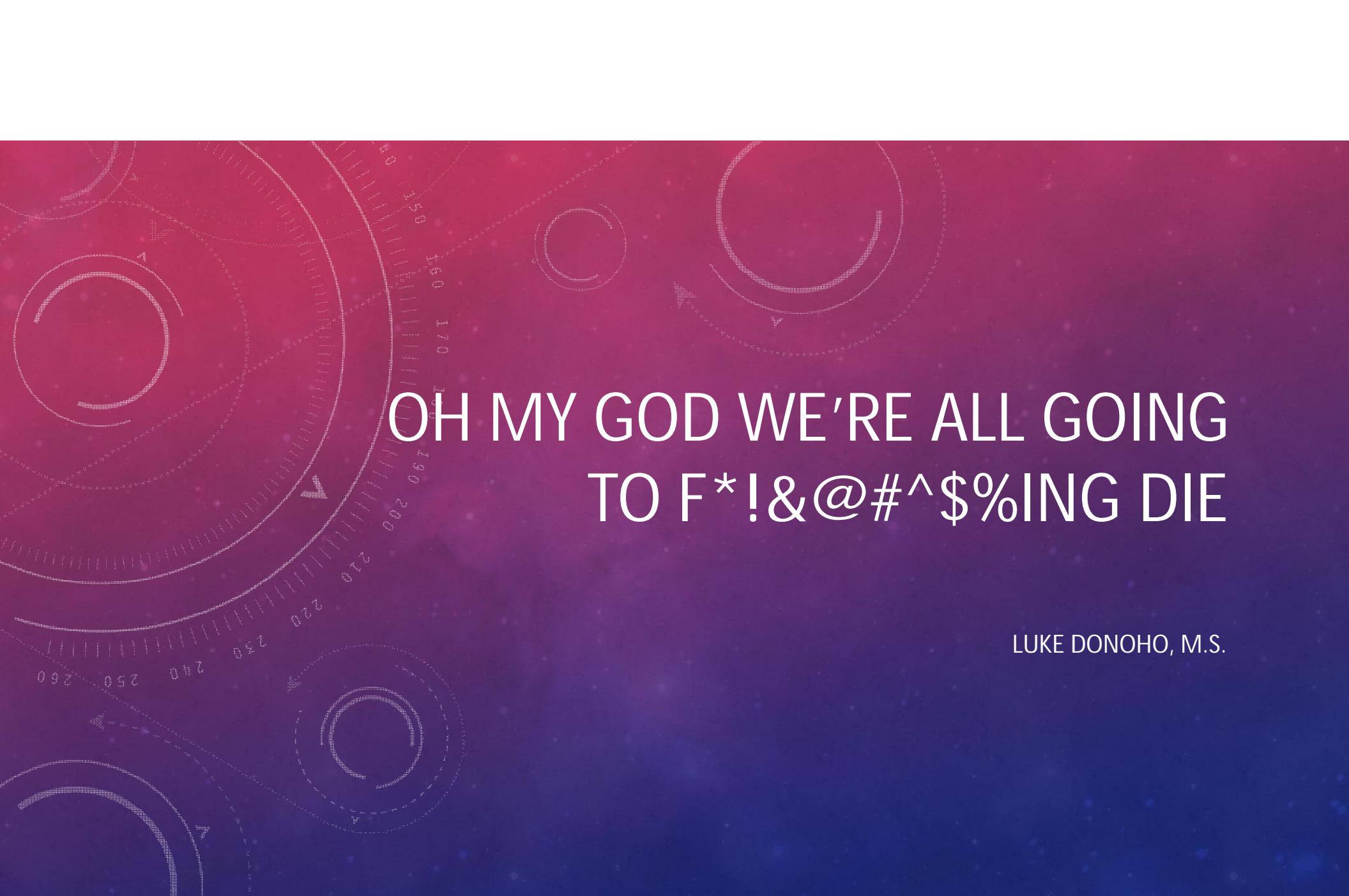




THREAT MODELING GLOBAL CATASTROPHIC RISKS

LUKE DONOHO, M.S.



OH MY GOD WE'RE ALL GOING
TO F*!&@#^\$%ING DIE

LUKE DONOHO, M.S.

AGENDA

- ❖ Existential Risk
- ❖ Threat Modeling
- ❖ Artificial Intelligence
- ❖ Nanotechnology
- ❖ Threat Modeling Existential Risk



EXISTENTIAL RISK

Future of Humanity Institute



Global Challenges Foundation

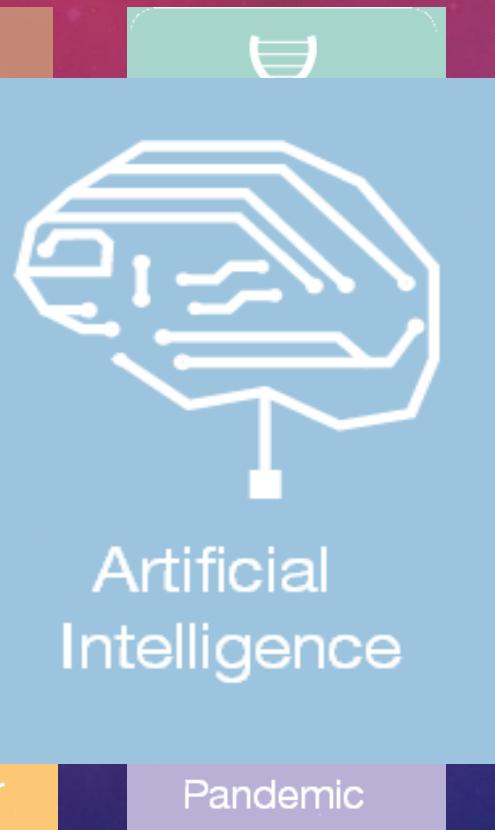


Global Challenges Foundation

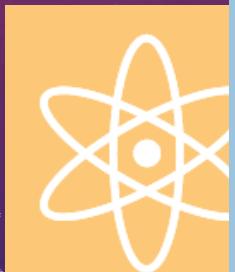
GLOBAL CATASTROPHIC RISKS



Super-volcano



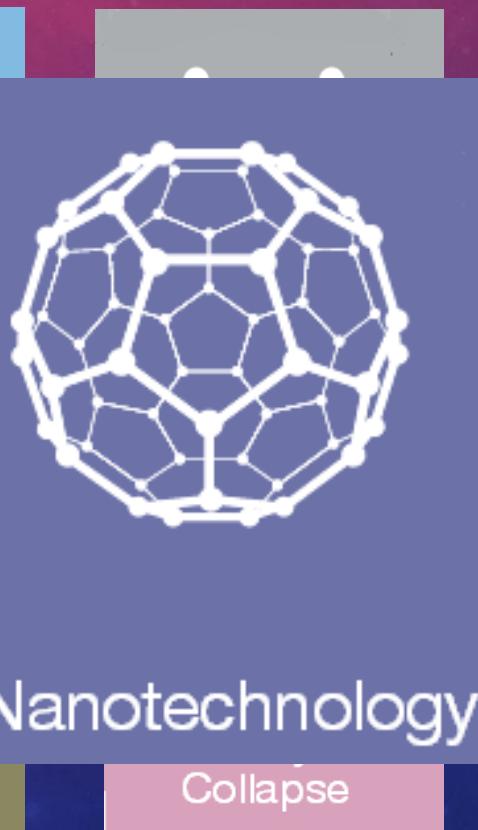
Artificial
Intelligence



Nuclear War



Major Asteroid Impact



Nanotechnology



Ecological
Catastrophe



Extreme
Climate Change



Unknown
Consequences

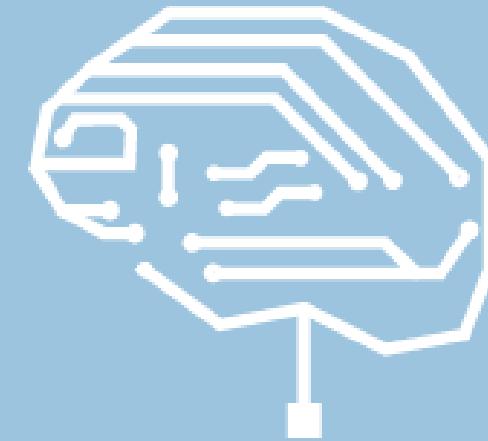
Reliability of AI Predictions

A single dominant AI or many

How intelligent AI will become

Whether extremely intelligent AI
can be controlled

Whether whole brain emulations
will come before AI



Artificial
Intelligence

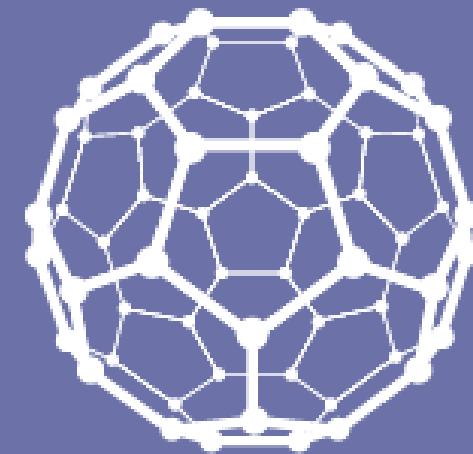
Timeline for development

Order of research progression

How quickly weapons can be assembled

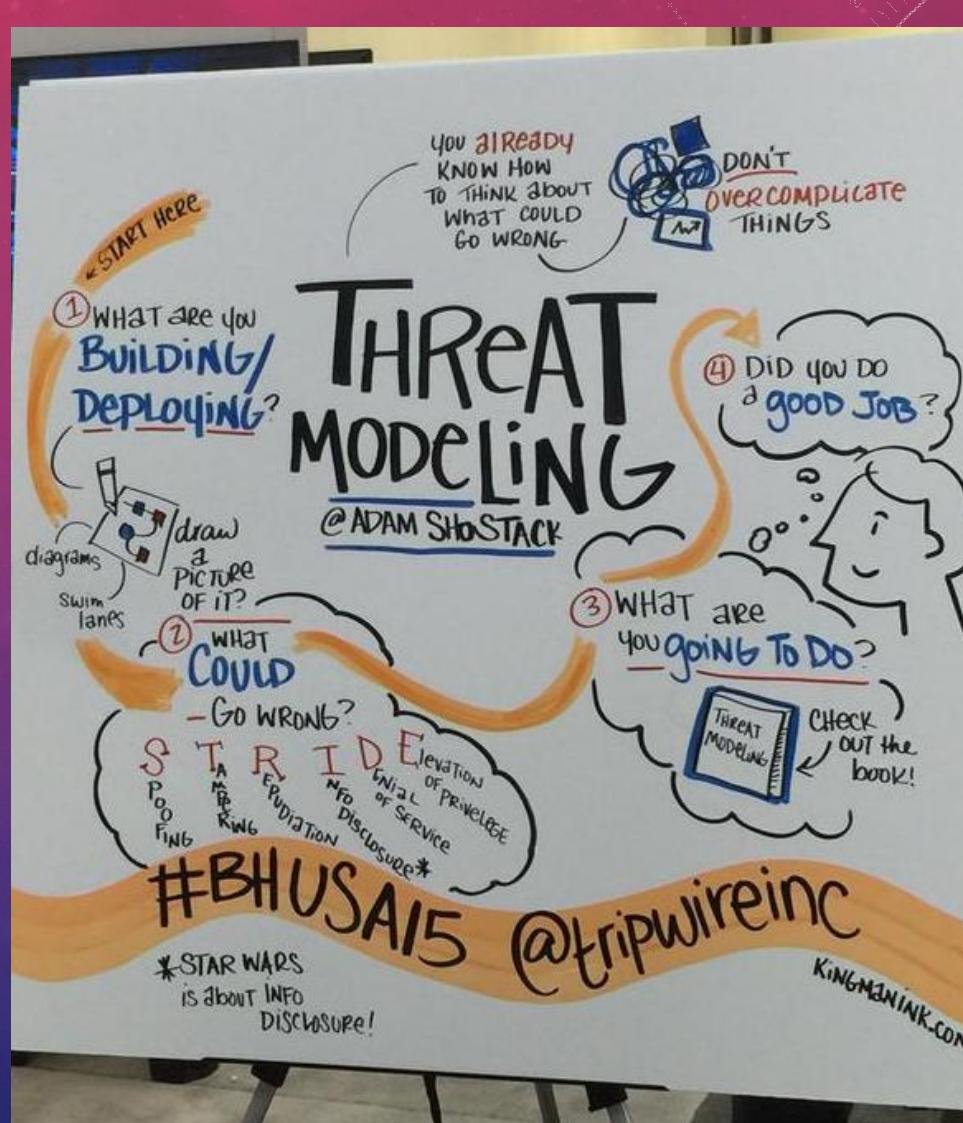
Surveillance capabilities of tools

Whether tools are made outside of human control

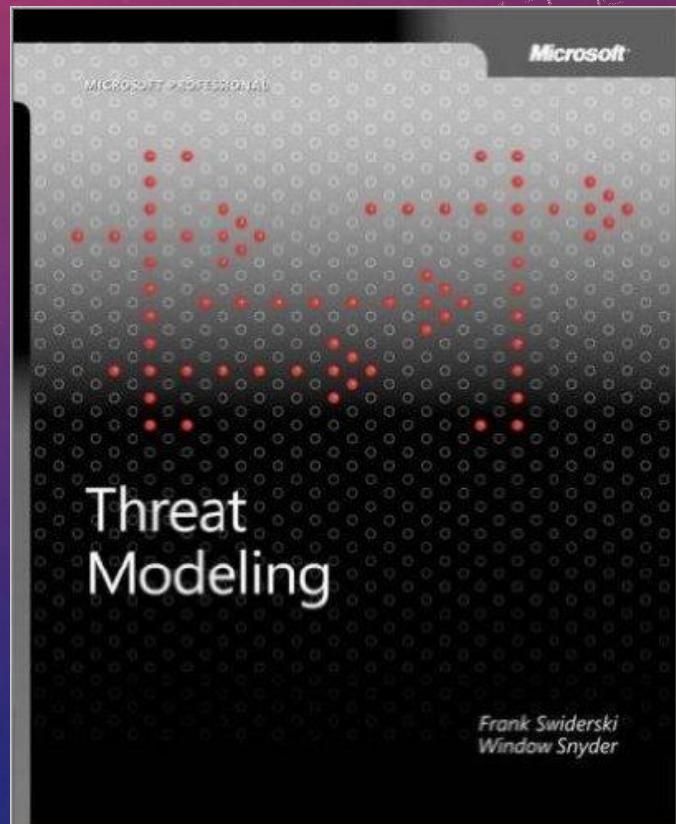


Nanotechnology

THREAT MODELING



THREAT MODELING – WINDOW SNYDER



“ Threat modeling is a methodology for identifying security vulnerabilities, for identifying the risks of a security vulnerability within that application ”

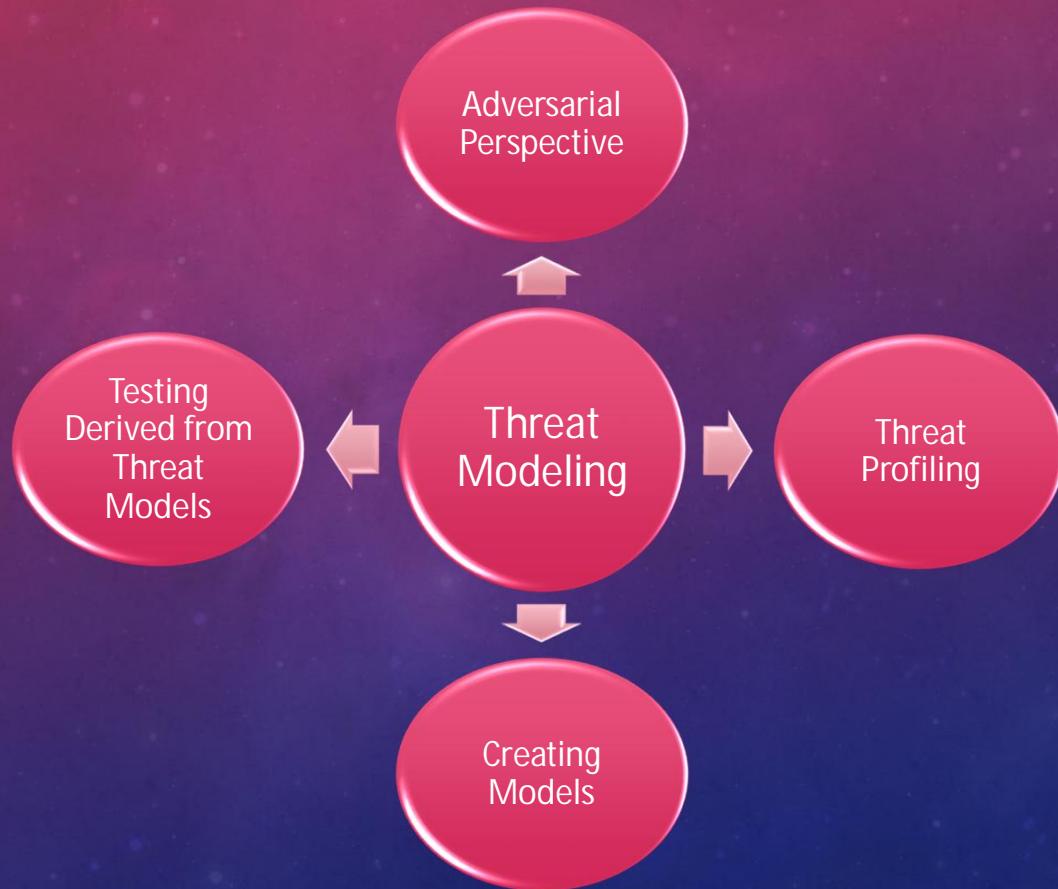
Window Snyder

THREAT MODELING – WINDOW SNYDER

fastly.



THREAT MODELING – WINDOW SNYDER

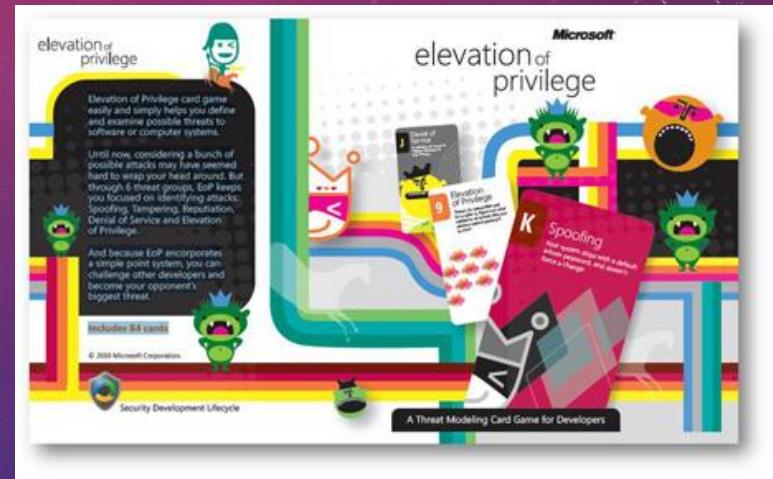


“Here's a pattern, and if we implement one architectural change we can eliminate all these vulnerabilities.”

Window Snyder

The goal, she said, is to remove whole categories of vulnerabilities.

THREAT MODELING – ADAM SHOSTACK

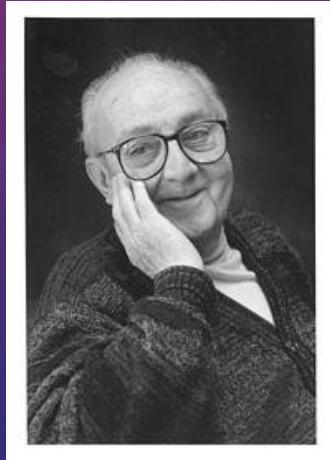


THREAT MODELING



<https://www.youtube.com/watch?v=KLpgoD8ySM>

“All models are wrong,
some models are useful”



- George Box

THREAT MODELING - STRIDE



THREAT MODELING – STRIDE SPOOFING

Impersonating something
or someone else



Spoofing

2. An attacker could squat on the random port or socket that the server normally uses
3. An attacker could try one credential after another and there's nothing to slow them down (online or offline)
4. An attacker can anonymously connect because we expect authentication to be done at a higher level
5. An attacker can confuse a client because there are too many ways to identify a server
6. An attacker can spoof a server because identifiers aren't stored on the client and checked for consistency on re-connection (that is, there's no key persistence)
7. An attacker can connect to a server or peer over a link that isn't authenticated (and encrypted)
8. An attacker could steal credentials stored on the server and reuse them (for example, a key is stored in a world readable file)
9. An attacker who gets a password can reuse it (Use stronger authenticators)

THREAT MODELING – STRIDE TAMPERING

Modifying Data or Code



Tampering

- 3. An attacker can take advantage of your custom key exchange or integrity control which you built instead
- 4. Your code makes access control decisions all over the
- 8. An attacker can manipulate data because there's no integrity protection for data on the network
- 9. An attacker can provide or control state information
- 10. An attacker can alter information in a data store because it has weak ACLs or includes a group which is equivalent to everyone ("all Live ID holders")
- J. An attacker can write to some resource because permissions are granted to the world or there are no ACLs continued on back

Tampering

THREAT MODELING – STRIDE REPUDIATION

Claiming to have not performed an action

R

Repudiation

2. An attacker can pass data through the log to attack a log reader, and there's no documentation of what sorts of validation are done
3. A low privilege attacker can read interesting security information in the logs
4. An attacker can alter digital signatures because the digital signature system you're implementing is weak, or uses MACs where it should use a signature
5. An attacker can alter log messages on a network because they lack strong integrity controls
6. An attacker can create a log entry without a timestamp (or no log entry is timestamped)
7. An attacker can make the logs wrap around and lose data
8. An attacker can make a log lose or confuse security information
9. An attacker can use a shared key to authenticate as different principals, confusing the information in the logs

THREAT MODELING – STRIDE INFORMATION DISCLOSURE

Exposing information to someone who should not have access

Information Disclosure

2. An attacker can brute-force file encryption because There's no defense in place (example defense: password Stretching)
3. An attacker can see error messages with security - Sensitive content
4. An attacker can read content because messages (say, An email or HTTP cookie) aren't encrypted even if the Channel is encrypted
8. An attacker can access information through a search indexer, logger, or other such mechanism
9. An attacker can read sensitive information in a file with bad ACLs
10. An attacker can read information in files with no ACLs

THREAT MODELING – STRIDE DENIAL OF SERVICE

Deny or Degrade
Services to Users



Denial of Service

2. An attacker can make your authentication system unusable or unavailable

3. An attacker can make a client unavailable or unusable

but the problem goes away when the attacker stops (client, authenticated, temporary)

4. An attacker can make a server unavailable or unusable but the problem goes away when the attacker stops (**server, authenticated, temporary**)

5. An attacker can make a client unavailable or unusable without ever authenticating but the problem goes away when the attacker stops (**client, anonymous, temporary**)

6. An attacker can make a server unavailable or unusable without ever authenticating but the problem goes away when the attacker stops (**server, anonymous, temporary**)

7. An attacker can make a client unavailable or unusable and the problem persists after the attacker goes away (**client, authenticated, persistent**)

THREAT MODELING – STRIDE ELEVATION OF PRIVILEGE

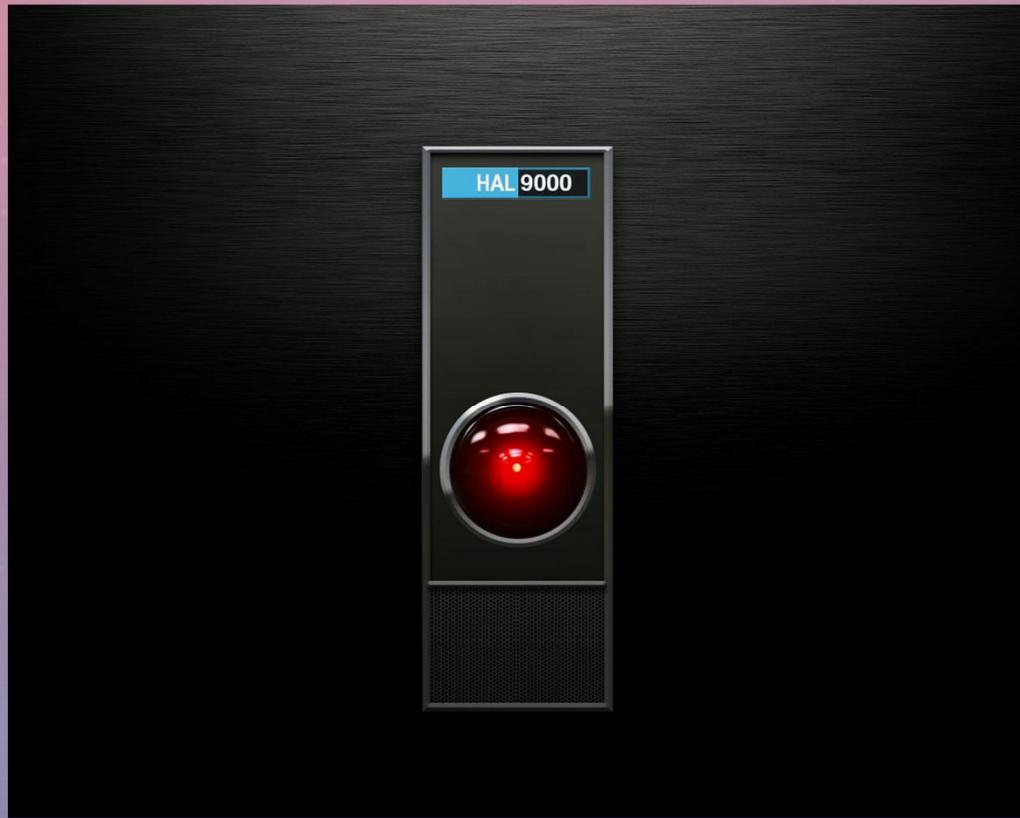
Gain capabilities without proper authorization



Elevation of Privilege (EoP)

- 5. An attacker can force data through different Validation paths which give different results
- 6. An attacker could take advantage of .NET permissions
- 7. An attacker can provide a pointer across a trust boundary. An attacker can reflect input back to a user, like cross site scripting
- Q. You include user-generated content within your page, possibly including the content of random URLs
- K. An attacker can inject a command that the system will run at a higher privilege level
- A. You've invented a new Elevation of Privilege attack

ARTIFICIAL INTELLIGENCE



ARTIFICIAL INTELLIGENCE

Overview

Ray Kurzweil

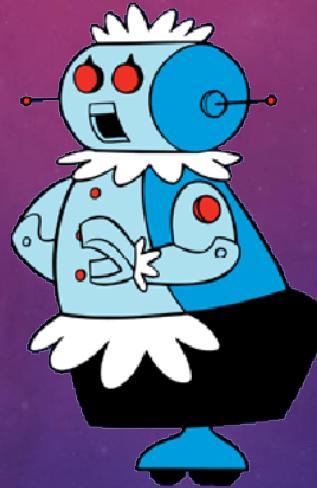
Eliezer Yudkowsky

Nick Bostrom

ARTIFICIAL INTELLIGENCE – WEAK AI



ARTIFICIAL INTELLIGENCE – STRONG AI (AGI)



ARTIFICIAL INTELLIGENCE – STRONG AI (AGI)



A



C



Look at the sample table of codes. The sample table lists codes for Department, Department Size, and Travel Budget. Below the table are two sample problems. In the problems, mark the circles below the codes that correspond to the information found in the table.

| Department | Department Size | Travel Budget |
|---------------------------|------------------------|------------------|
| 31 Marketing | 1 1-9 employees | A \$500-999 |
| 25 Operations | 2 10-19 employees | B \$1,000-2,499 |
| 42 Accounting | 3 20-39 employees | C \$2,500-5,999 |
| 11 Research & Development | 4 40-59 employees | D \$6,000-7,999 |
| 28 Personnel | 5 60 or more employees | E \$8,000-10,000 |

S1. Accounting Dept., staff of 10, \$3,000 Travel Budget.

Department Department Size Travel Budget

11 25 28 31 42

1 2 3 4 5

A B C D E

S1. Marketing Dept., staff of 45, \$9,500 Travel Budget.

Department Department Size Travel Budget

11 25 28 31 42

1 2 3 4 5

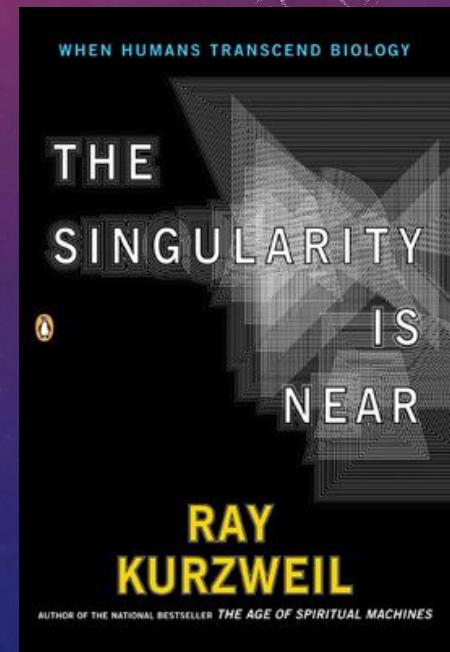
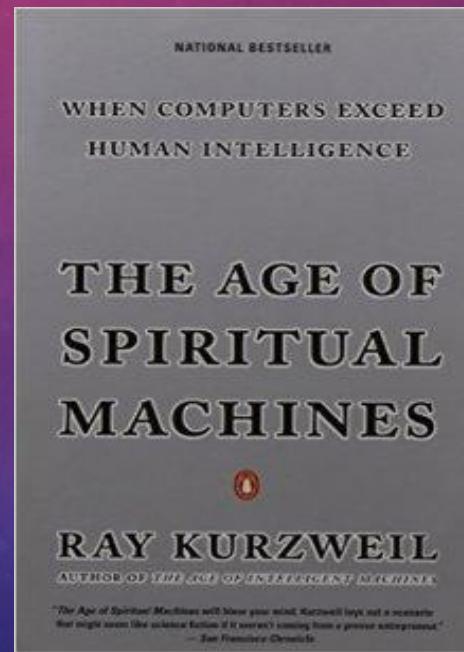
A B C D E



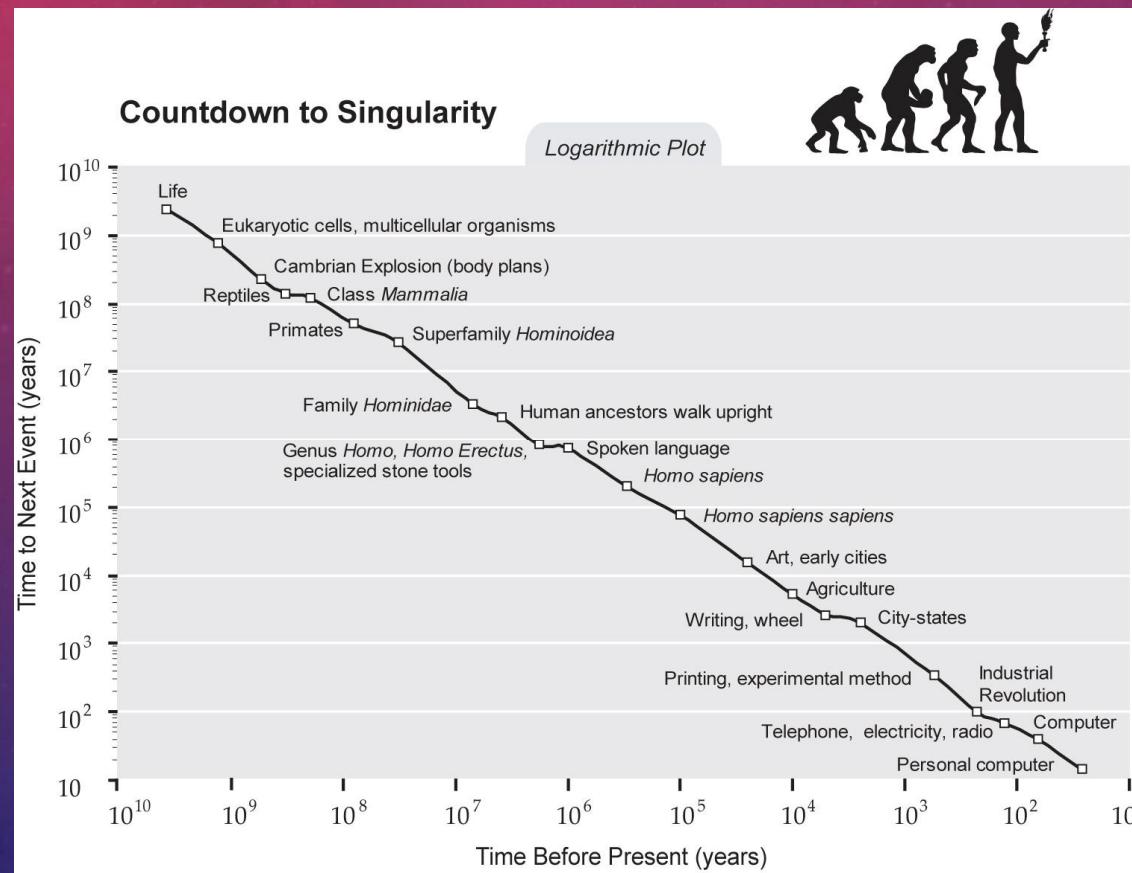
ARTIFICIAL INTELLIGENCE – SUPERINTELLIGENCE



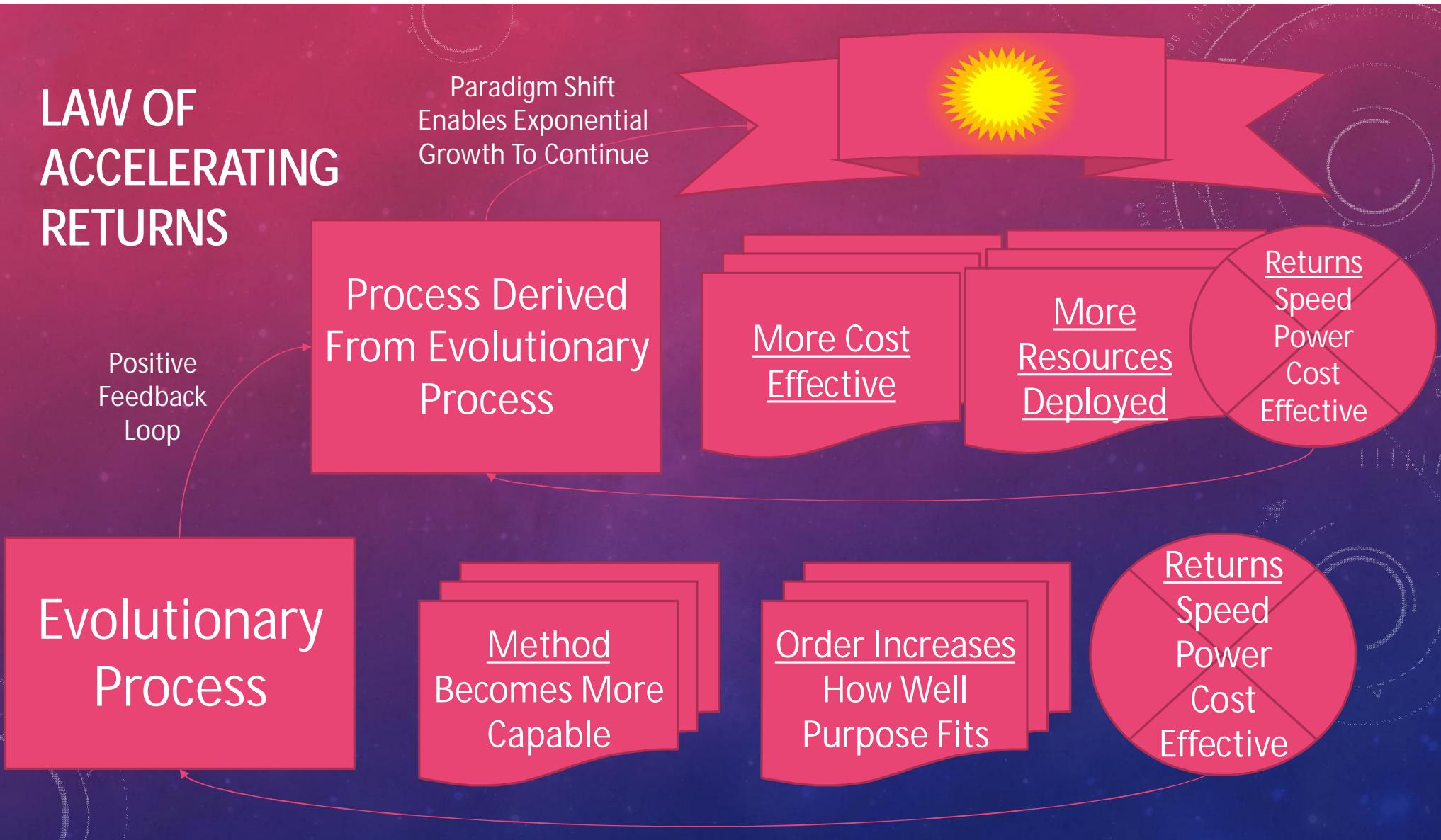
ARTIFICIAL INTELLIGENCE – RAY KURZWEIL



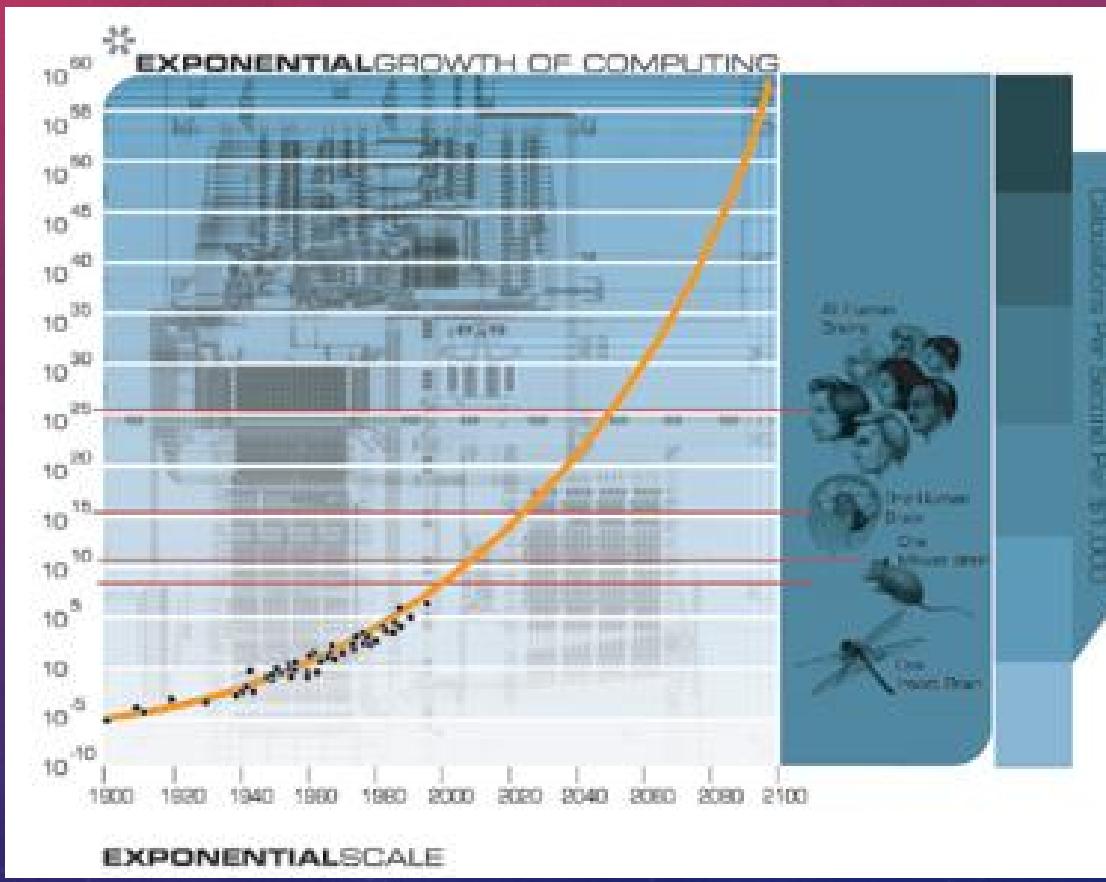
RAY KURZWEIL - SINGULARITY



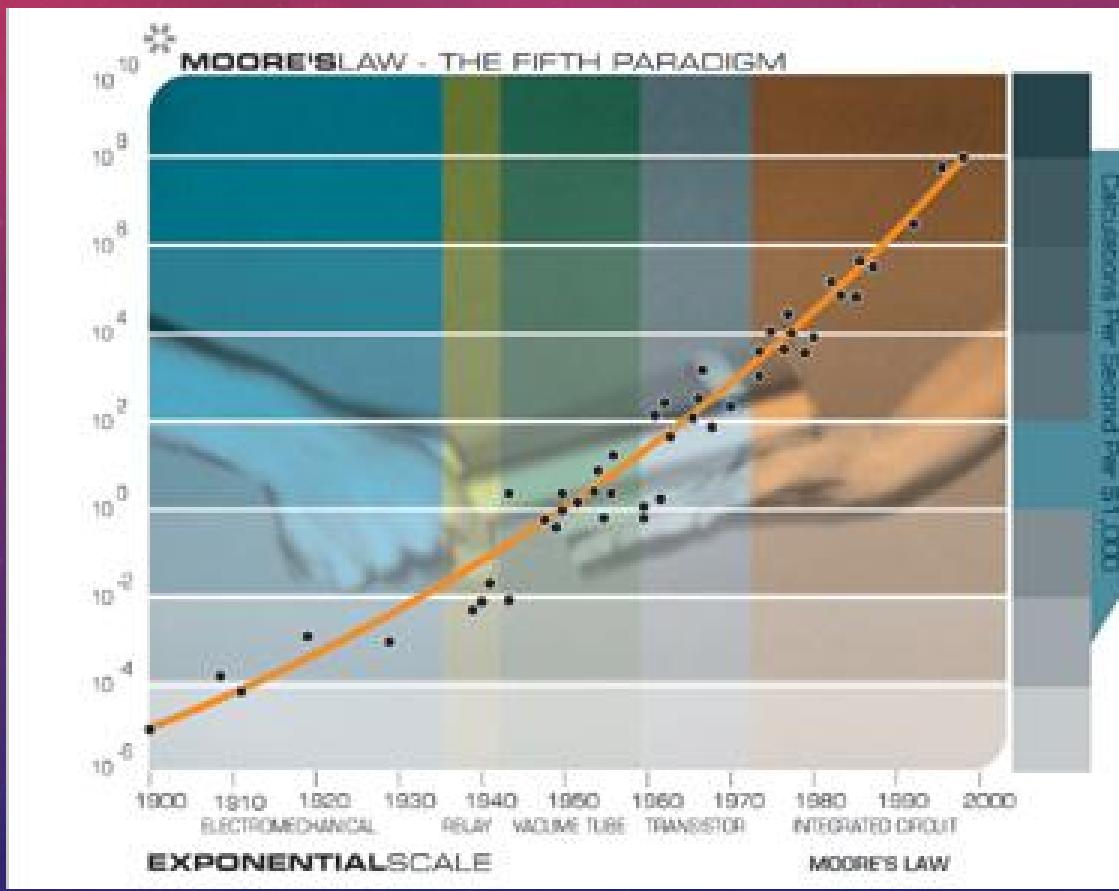
LAW OF ACCELERATING RETURNS



RAY KURZWEIL – LAW OF ACCELERATING RETURNS



RAY KURZWEIL – MOORE'S LAW



ARTIFICIAL INTELLIGENCE – ELIEZER YUDKOWSKY



MIRI
MACHINE INTELLIGENCE
RESEARCH INSTITUTE

ARTIFICIAL INTELLIGENCE – ELIEZER YUDKOWSKY



ARTIFICIAL INTELLIGENCE – COHERENT EXTRAPOLATED VOLITION (CEV)

Knew More

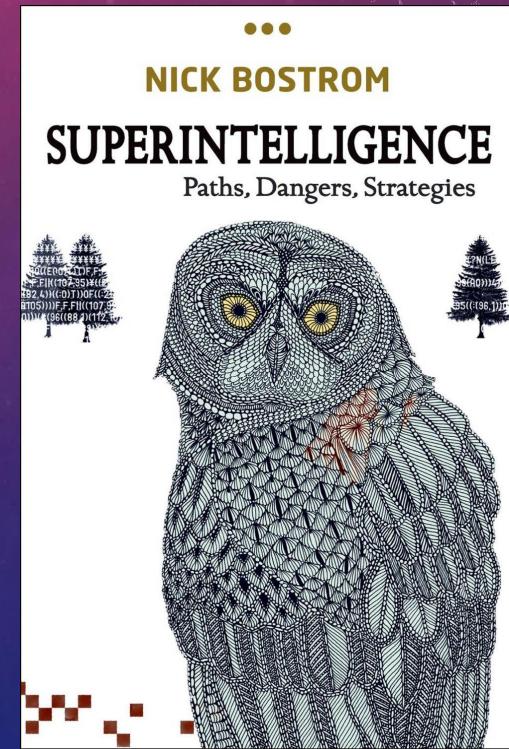
Thought Faster

CEV

Were More The People
We Wish We Were

Had Grown Up Farther
Together

ARTIFICIAL INTELLIGENCE – NICK BOSTROM



NICK BOSTROM – EXISTENTIAL RISK CLASSIFICATION



Extinction



Human / No Posthuman



Undesirable



Unrealized Potential

NICK BOSTROM – EXISTENTIAL RISK CLASSIFICATION

Bangs

- Misuse of Nanotechnology
- Nuclear Holocaust
- Simulation
- Genetic Engineering
- Nanotechnology
- Accident
- Physics Disaster
- Naturally Occurring Disease
- Asteroid / Comet
- Global Warming

Crunches

- Resource Depletion
- Misguided World Government
- Technological Arrest

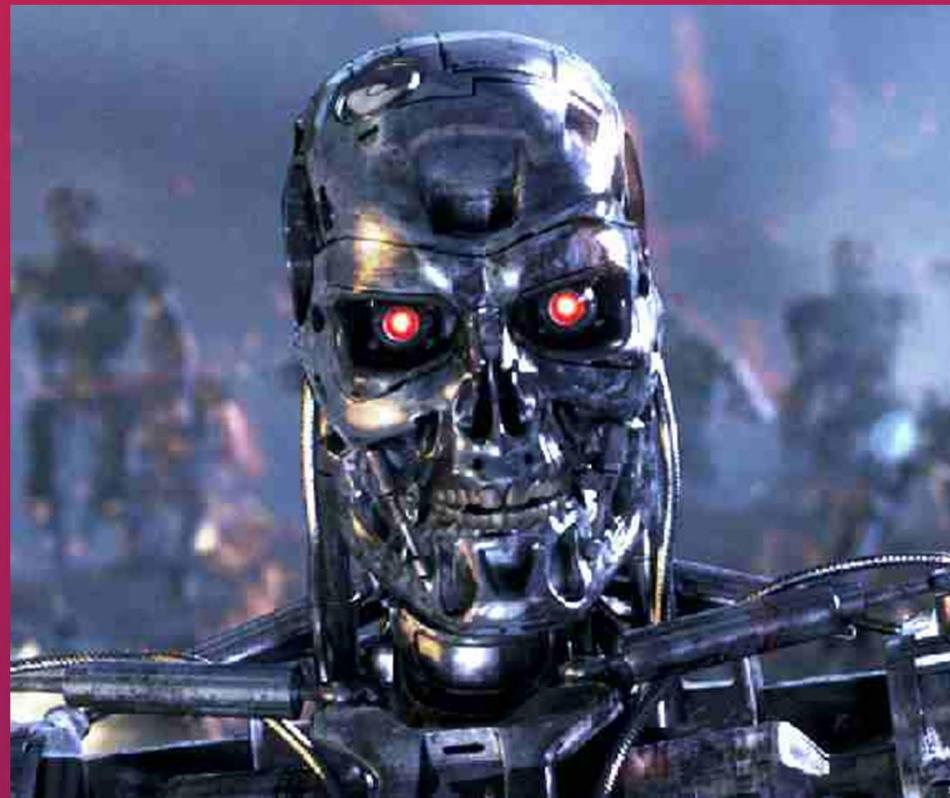
Shrieks

- Transcending Upload Takeover
- Flawed ASI
- Oppressive Regime

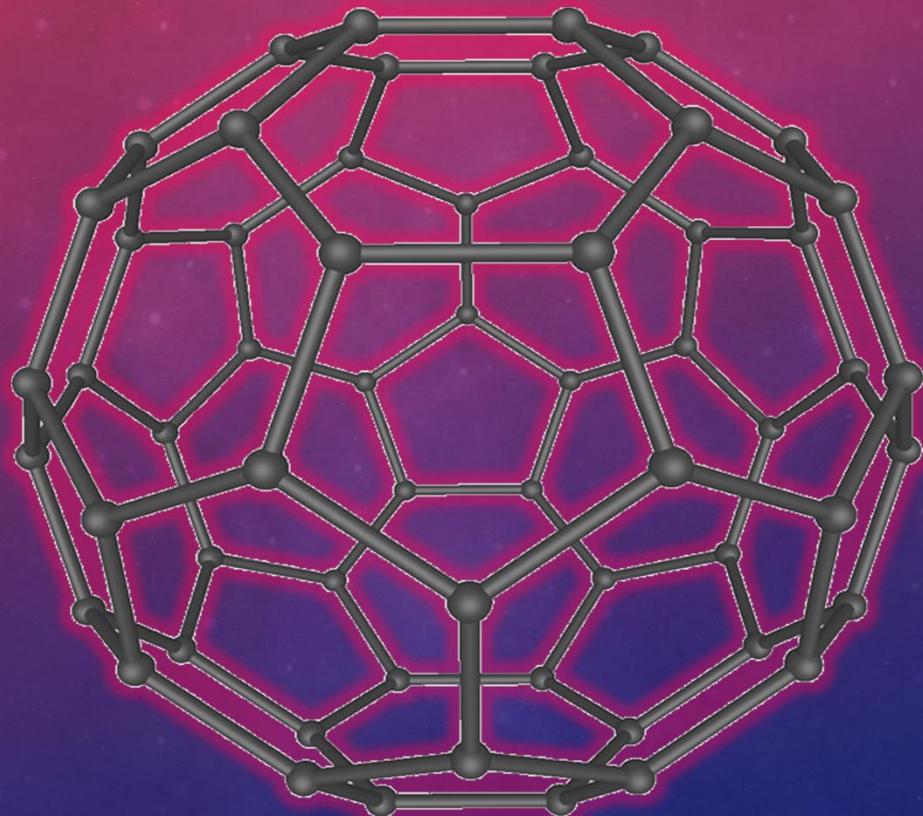
Whimpers

- Potential Eroded By Evolution
- Killed By Aliens

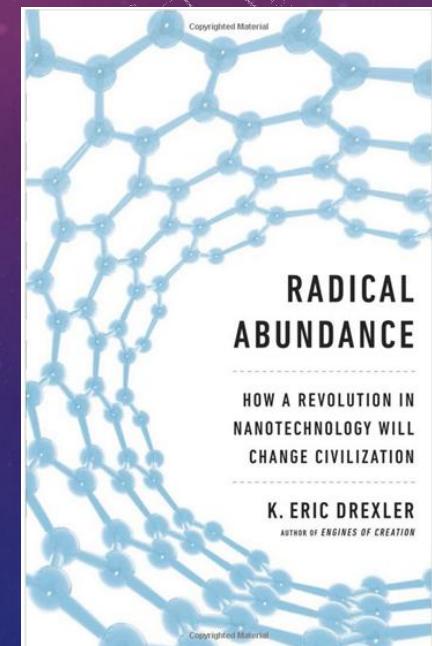
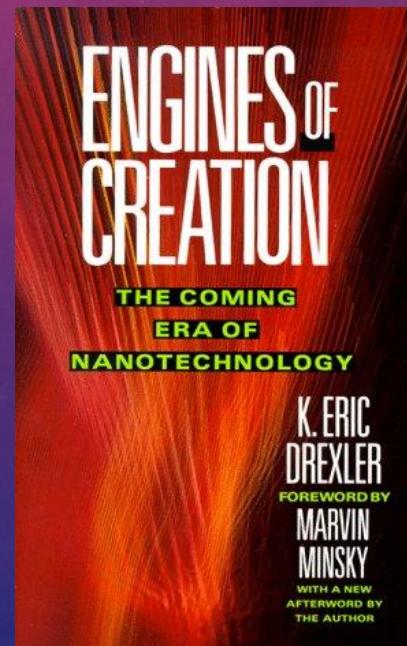
NICK BOSTRÖM - RECALCITRANCE



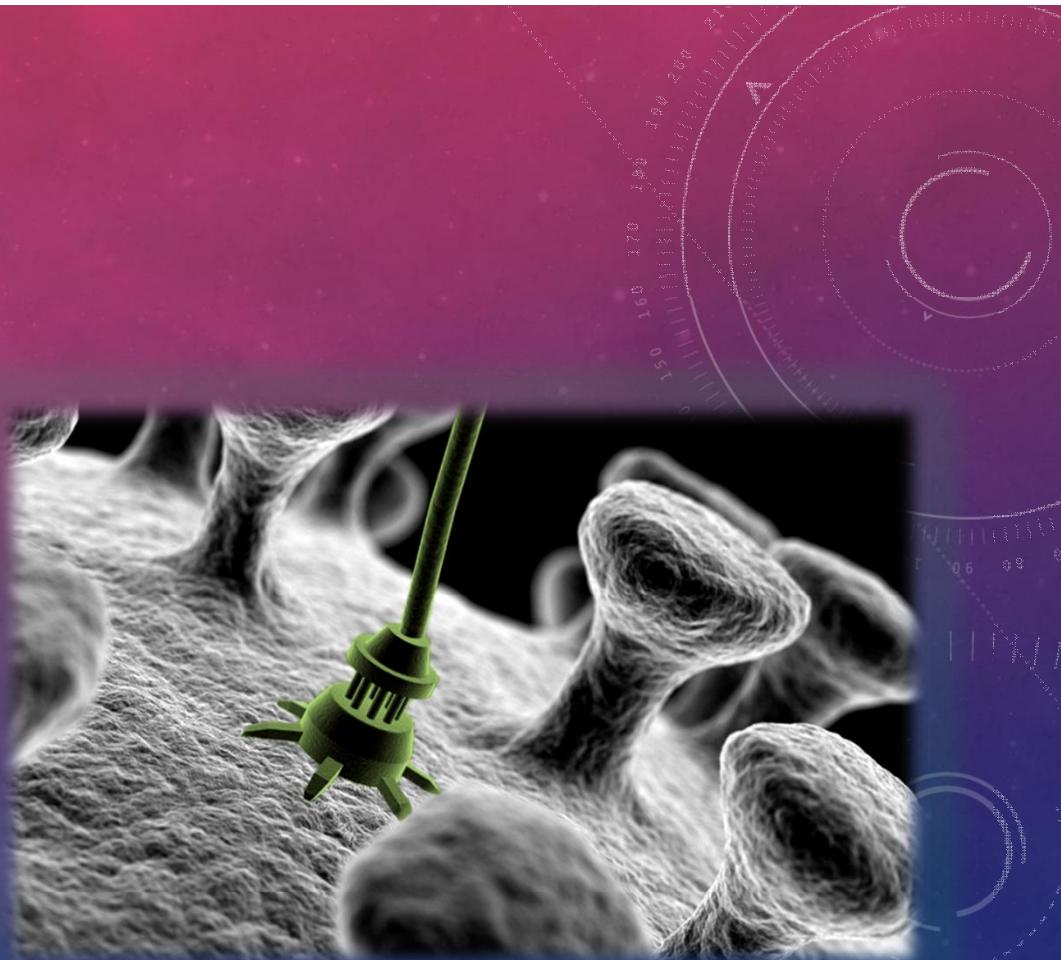
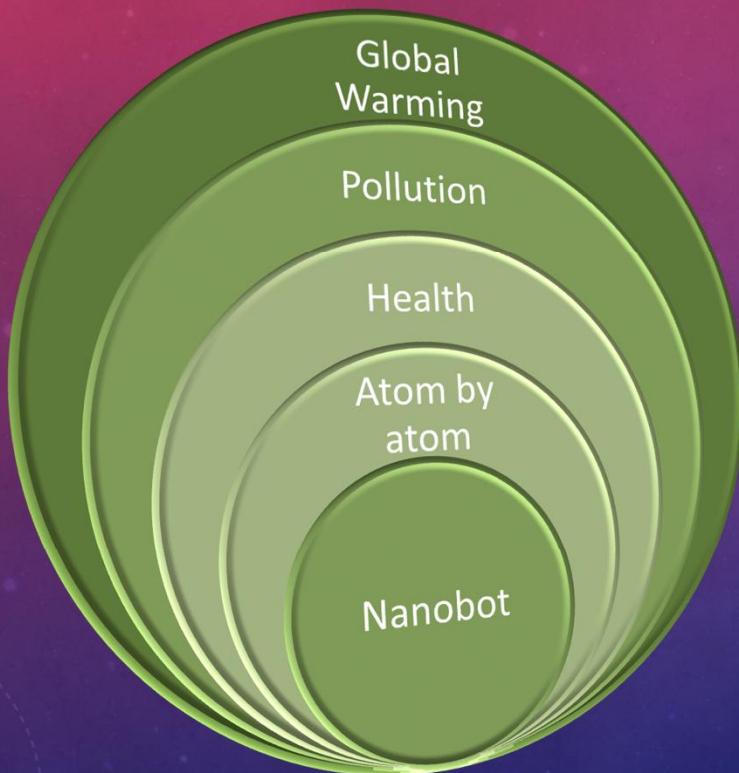
NANOTECHNOLOGY



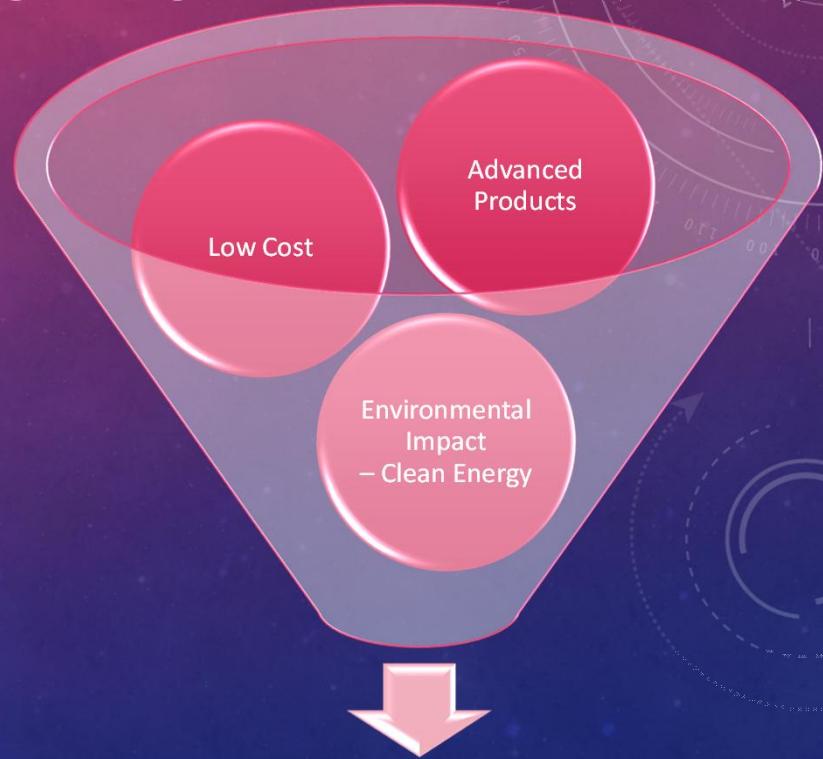
NANOTECHNOLOGY – K. ERIC DREXLER



K. ERIC DREXLER ENGINES OF CREATION



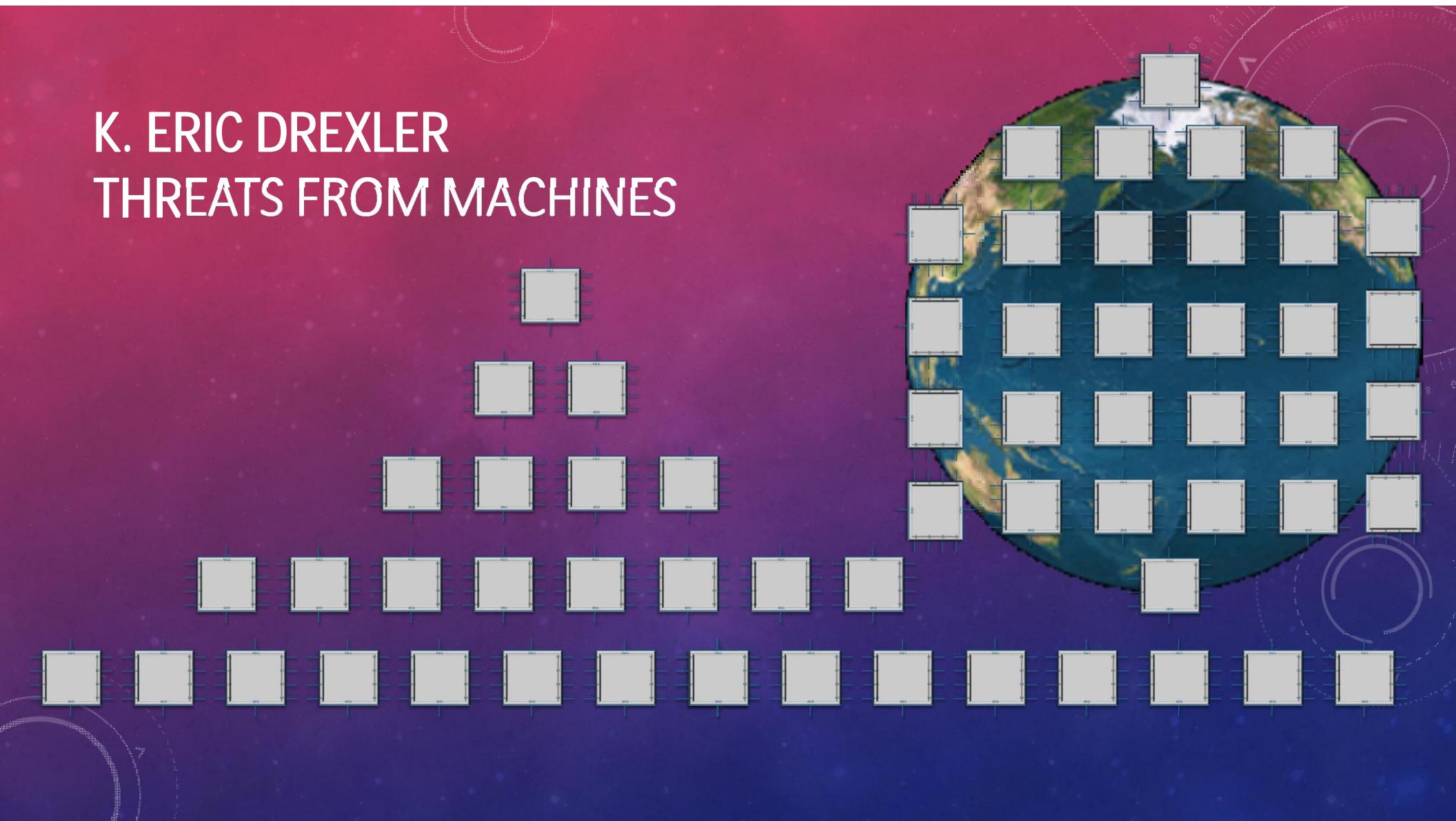
K. ERIC DREXLER ATOMICALLY PRECISE MANUFACTURING



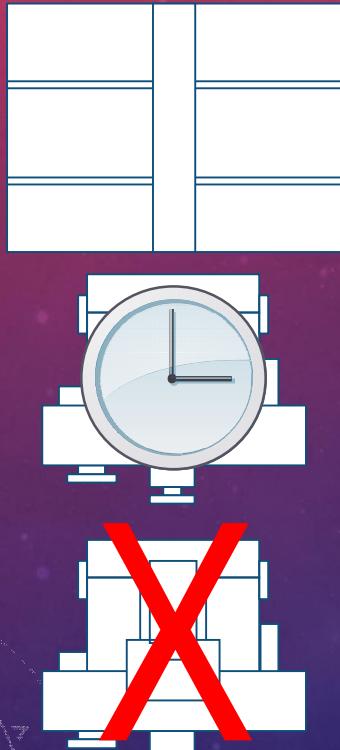
End of Industrial Revolution?

K. ERIC DREXLER

THREATS FROM MACHINES



K. ERIC DREXLER THREATS FROM ASSEMBLER BREAKTHROUGH



K. ERIC DREXLER PROBABILITY OF SUCCESS

Success is probable based on our experiences

Replicators are physical systems

Biological life has been defending against invasive replicators since creation

AI can become a Nanomanager

AI – SW

Nanotechnology - HW

THREAT MODELING EXISTENTIAL RISK: NICK BOSTROM - SUPERINTELLIGENCE



<https://github.com/hello7s/Existential-Threat-Models>