AI for Good

Global Summit

Rethinking Security in the Age of LLMs and Agentic AI

Sounil Yu
CTO, Knostic

11 July 2025
Geneva, Switzerland

$whoami

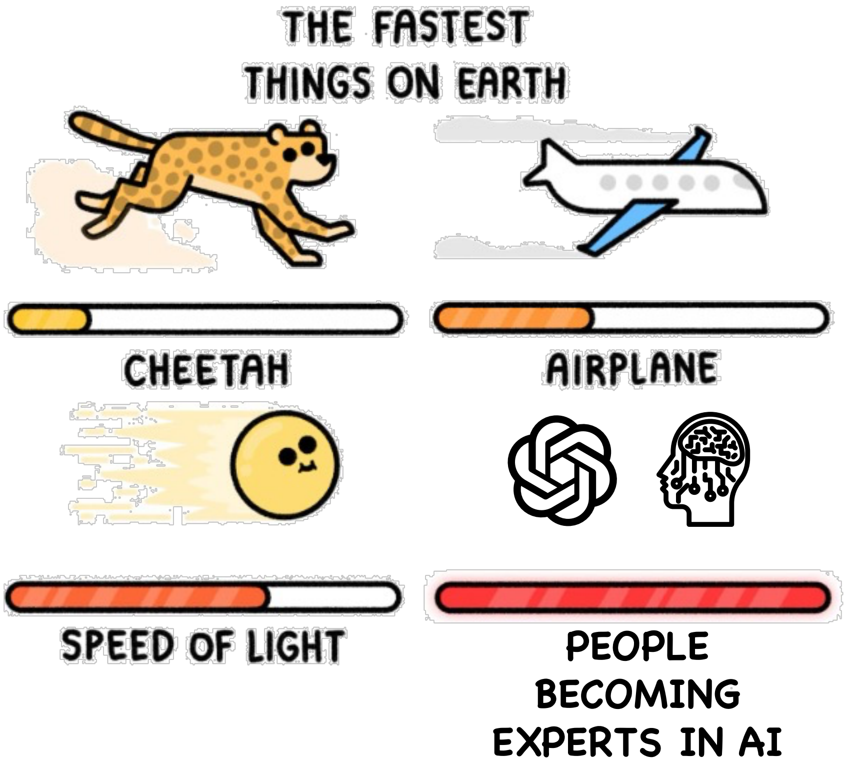
Currently: Chief AI Safety Officer/Co-Founder at KNOSTIC

Formerly: Chief Security Scientist at BANK OF AMERICA

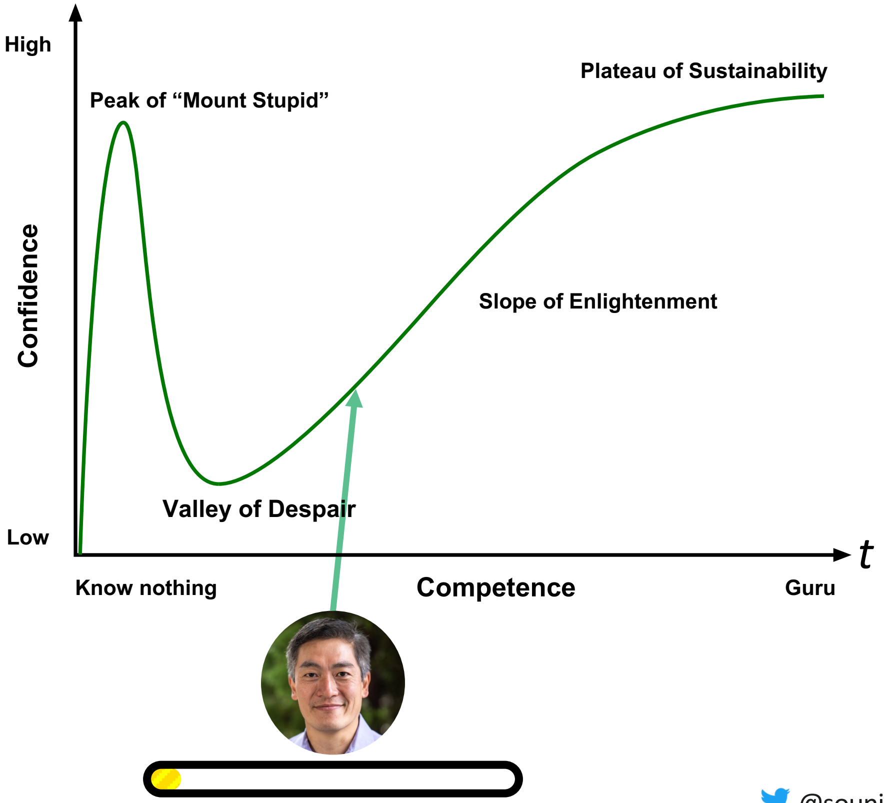
Other Random Recognitions:

- Lifetime Achievement Award
- Cybersecurity Hall of Fame
- Most Influential Person in Security
- Cyberscoop Cybersecurity Visionary
- 20+ Patents

$whoaiml

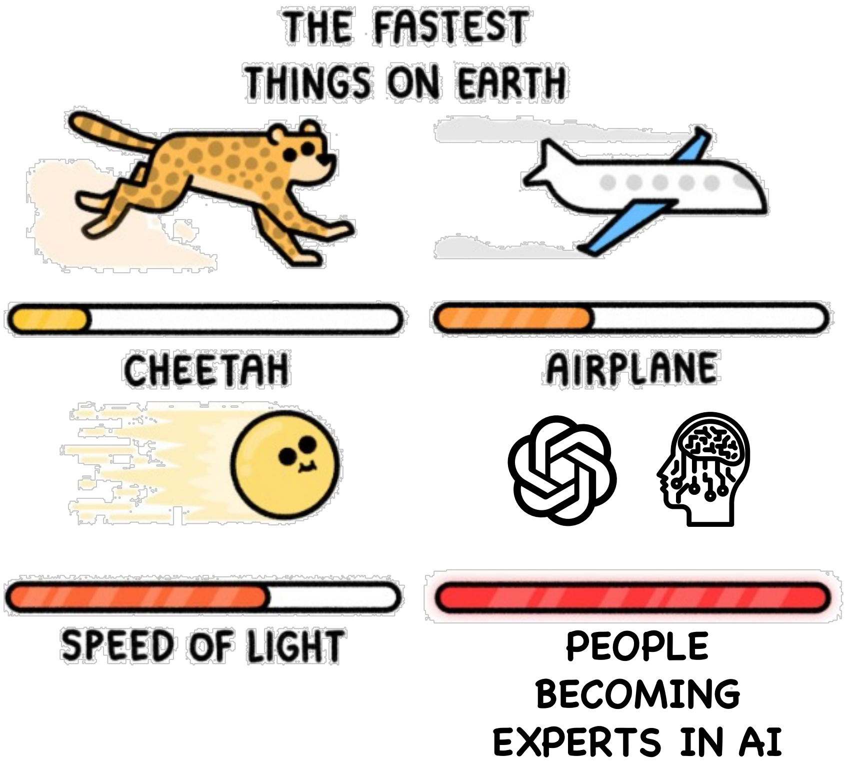


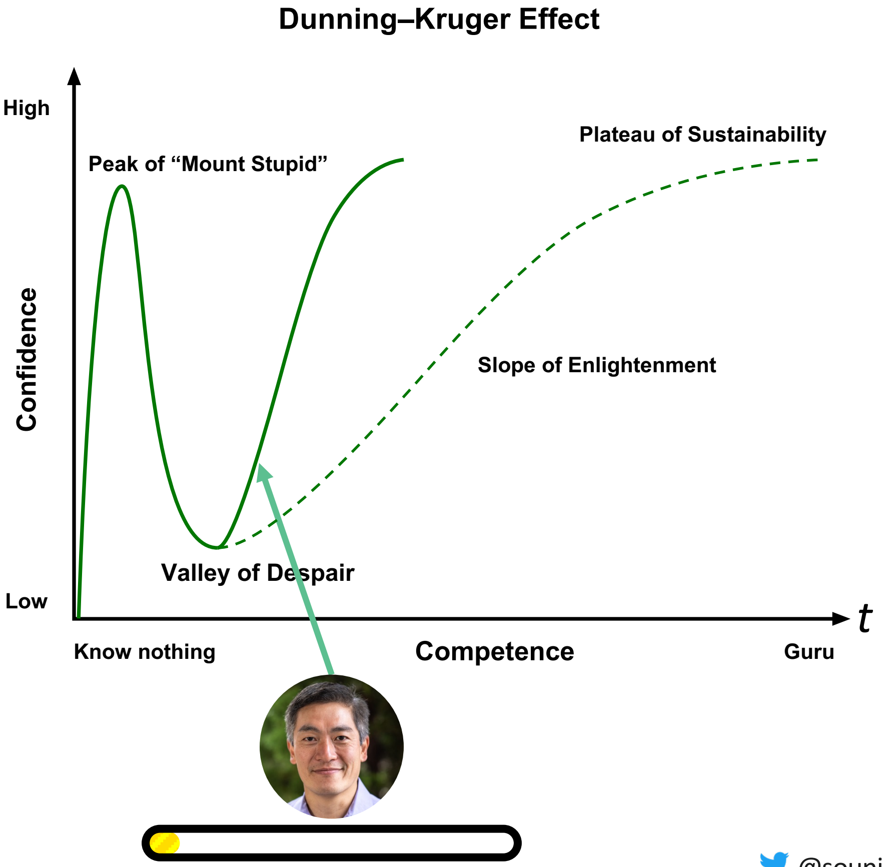
Dunning–Kruger Effect



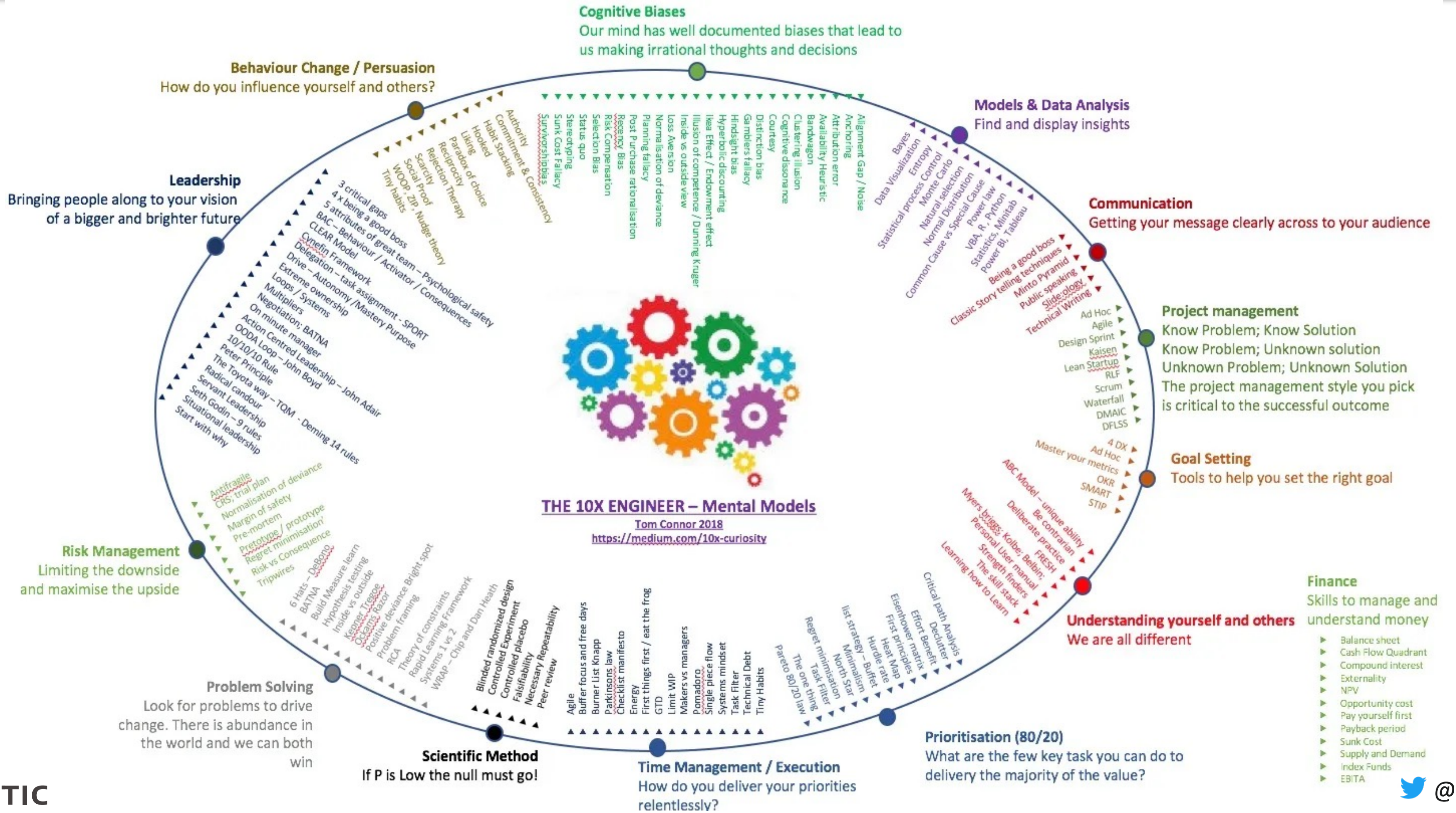
How do you accelerate competence?

AI for Good
Global Summit
11 July 2025
Geneva, Switzerland





Accelerating Competence w/ Mental Models



THE 10X ENGINEER – Mental Models

Tom Connor 2018

https://medium.com/10x-curiosity

Cognitive Biases
Our mind has well documented biases that lead to us making irrational thoughts and decisions

- Alignment Gap / Noise
- Anchoring
- Attribution error
- Availability Heuristic
- Bandwagon
- Clustering illusion
- Cognitive dissonance
- Courtesy
- Distinction bias
- Gamblers fallacy
- Hindsight bias
- Hyperbolic discounting
- Ikea Effect / Endowment effect
- Illusion of competence / Dunning Kruger
- Inside vs outside view
- Loss Aversion
- Normalisation of deviance
- Planning fallacy
- Post Purchase rationalisation
- Recency Bias
- Risk Compensation
- Selection Bias
- Status quo
- Stereotyping
- Sunk Cost Fallacy
- Survivorshipbias

Models & Data Analysis
Find and display insights

- Bayes
- Data Visualization
- Entropy
- Statistical process Control
- Monte Carlo
- Natural selection
- Normal Distribution
- Common Cause vs Special Cause
- Power law
- VBA, R, Python
- Statistics, Minitab
- Power BI, Tableau

Communication
Getting your message clearly across to your audience

- Being a good boss
- Classic Story telling techniques
- Minto Pyramid
- Public speaking
- Slide:ology
- Technical Writing

Project management
Know Problem; Know Solution
Know Problem; Unknown solution
Unknown Problem; Unknown Solution
The project management style you pick is critical to the successful outcome

- Ad Hoc
- Agile
- Design Sprint
- Kaisen
- Lean Startup
- RLF
- Scrum
- Waterfall
- DMAIC
- DFLSS

Goal Setting
Tools to help you set the right goal

- 4 DX
- Ad Hoc
- Master your metrics
- OKR
- SMART
- STIP

Understanding yourself and others
We are all different

- ABC Model – unique ability
- Be contrarian
- Deliberate practice
- FRESH
- Myers briggs; Kolbe; Belbin;
- Personal User manual
- Strength finders
- The skill stack
- Learning how to Learn

Finance
Skills to manage and understand money

- Balance sheet
- Cash Flow Quadrant
- Compound interest
- Externality
- NPV
- Opportunity cost
- Pay yourself first
- Payback period
- Sunk Cost
- Supply and Demand
- Index Funds
- EBITA

Prioritisation (80/20)
What are the few key task you can do to delivery the majority of the value?

- Critical path Analysis
- Declutter
- Effort Benefit
- Eisenhower matrix
- First principles
- Heat Map
- Hurdle rate
- list strategy – Buffet
- Minimalism
- North Star
- Regret minimisation
- Task Filter
- The one thing
- Pareto 80/20 law

Time Management / Execution
How do you deliver your priorities relentlessly?

- Agile
- Buffer focus and free days
- Burner List Knapp
- Parkinsons law
- Checklist manifesto
- Energy
- First things first / eat the frog
- GTD
- Limit WIP
- Makers vs managers
- Pomadoro
- Single piece flow
- Systems mindset
- Task Filter
- Technical Debt
- Tiny Habits

Scientific Method
If P is Low the null must go!

- Blinded randomized design
- Controlled Experiment
- Controlled placebo
- Falsifiability
- Necessary Repeatability
- Peer review

Problem Solving
Look for problems to drive change. There is abundance in the world and we can both win

- 6 Hats – DeBono
- BATNA
- Build Measure learn
- Hypothesis testing
- Inside vs outside
- Kepner Tregoe
- Ockams Razor
- Positive deviance Bright spot
- Problem framing
- RCA
- Theory of constraints
- Rapid Learning Framework
- Systems 1 vs 2
- WRAP – Chip and Dan Heath

Risk Management
Limiting the downside and maximise the upside

- Antifragile
- CRS; trial plan
- Normalisation of deviance
- Margin of safety
- Pre-mortem
- Pretotype / prototype
- Regret minimisation'
- Risk vs Consequence
- Tripwires

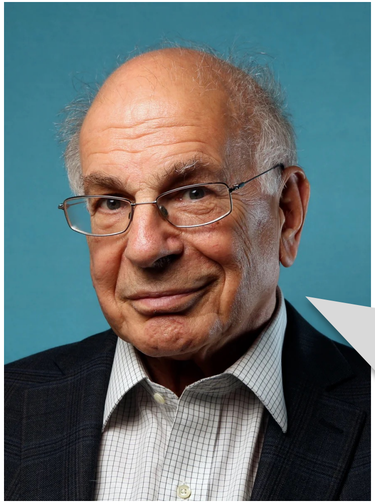
Leadership
Bringing people along to your vision of a bigger and brighter future

- 3 critical gaps
- 4 x being a good boss
- 5 attributes of great team – Psychological safety
- BAC – Behaviour / Activator / Consequences
- CLEAR Model
- Cynefin Framework
- Delegation – task assignment - SPORT
- Drive – Autonomy /Mastery Purpose
- Extreme ownership
- Loops / Systems
- Multipliers
- Negotiation; BATNA
- On minute manager
- Action Centred Leadership – John Adair
- OODA Loop – John Boyd
- 10/10/10 Rule
- Peter Principle
- The Toyota way – TQM - Deming 14 rules
- Radical candour
- Servant Leadership
- Seth Godin – 9 rules
- Situational leadership
- Start with why

Behaviour Change / Persuasion
How do you influence yourself and others?

- Authority
- Commitment & Consistency
- Habit Stacking
- Hooked
- Liking
- Paradox of choice
- Reciprocity
- Rejection Therapy
- Scarcity
- Social Proof
- WOOP, ZIP, Nudge theory
- Tiny habits

Two Caveats

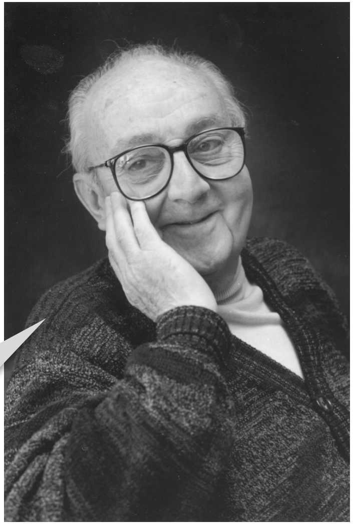


Daniel Kahneman

Theory induced blindness:

Once you have accepted a theory and used it as a tool in your thinking, it is extraordinarily difficult to notice its flaws

George E. P. Box



All models are wrong... but some are useful

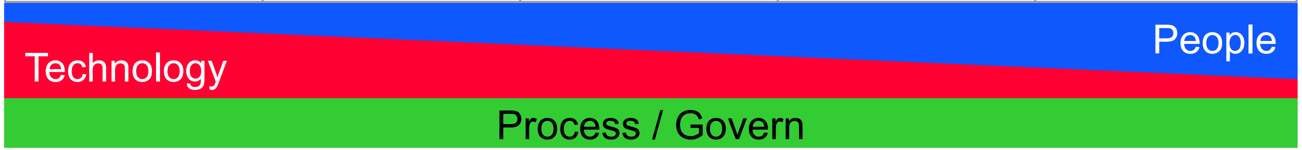
The Cyber Defense Matrix

AI for Good
Global Summit
11 July 2025
Geneva, Switzerland

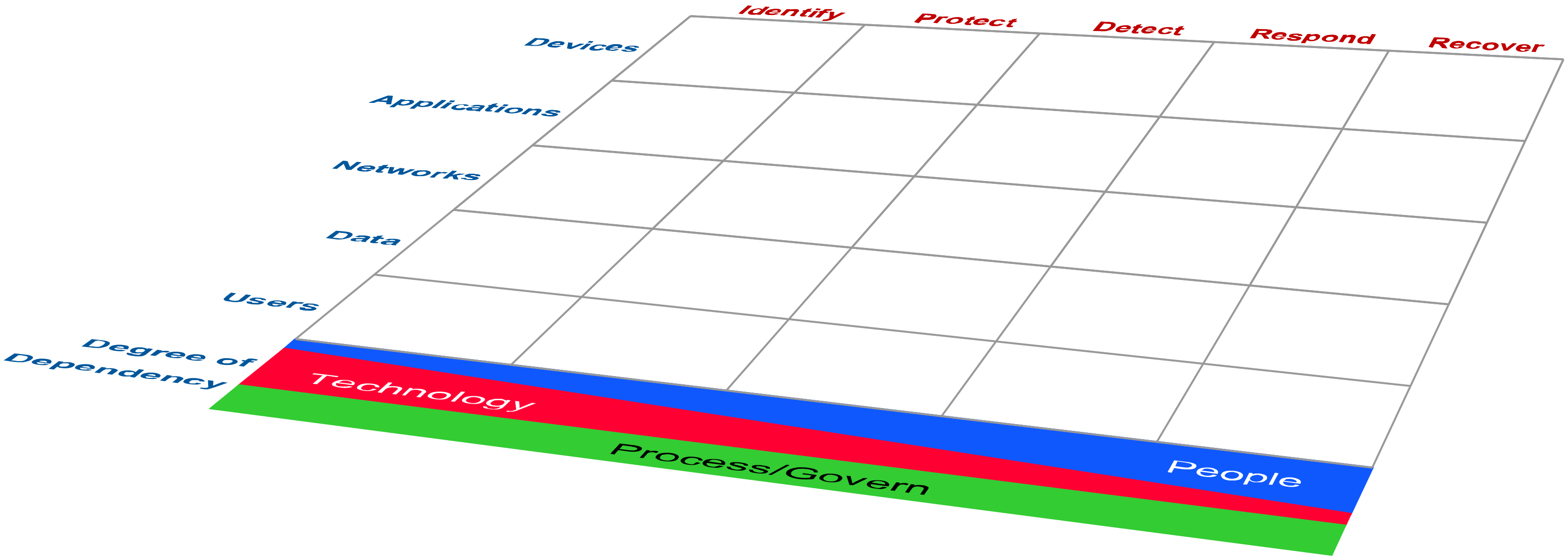
	Identify	Protect	Detect	Respond	Recover
Devices					
Applications					
Networks					
Data					
Users					

Where does AI fit?

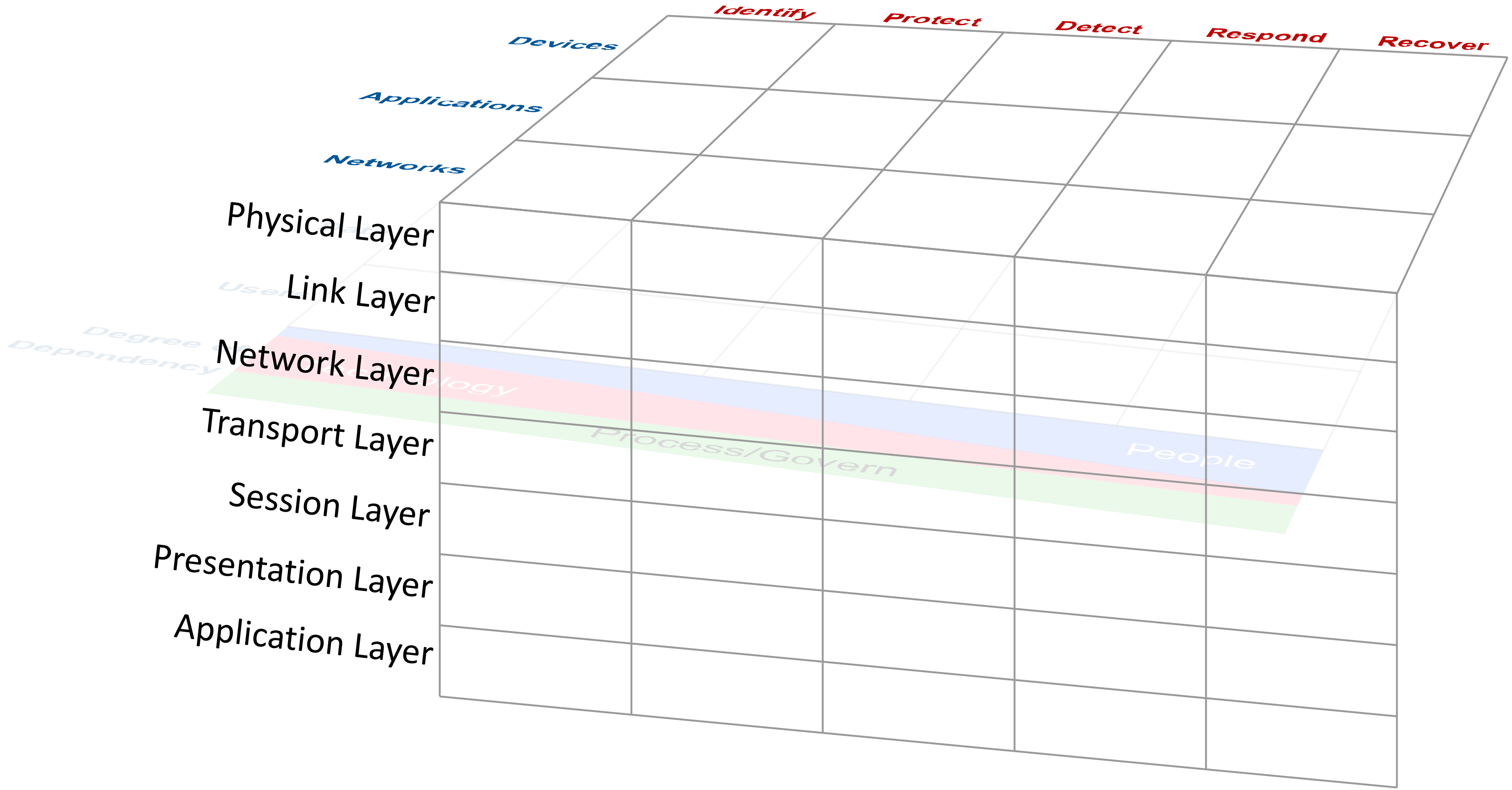
Degree of Dependency



Cyber Defense Matrix in 3D!

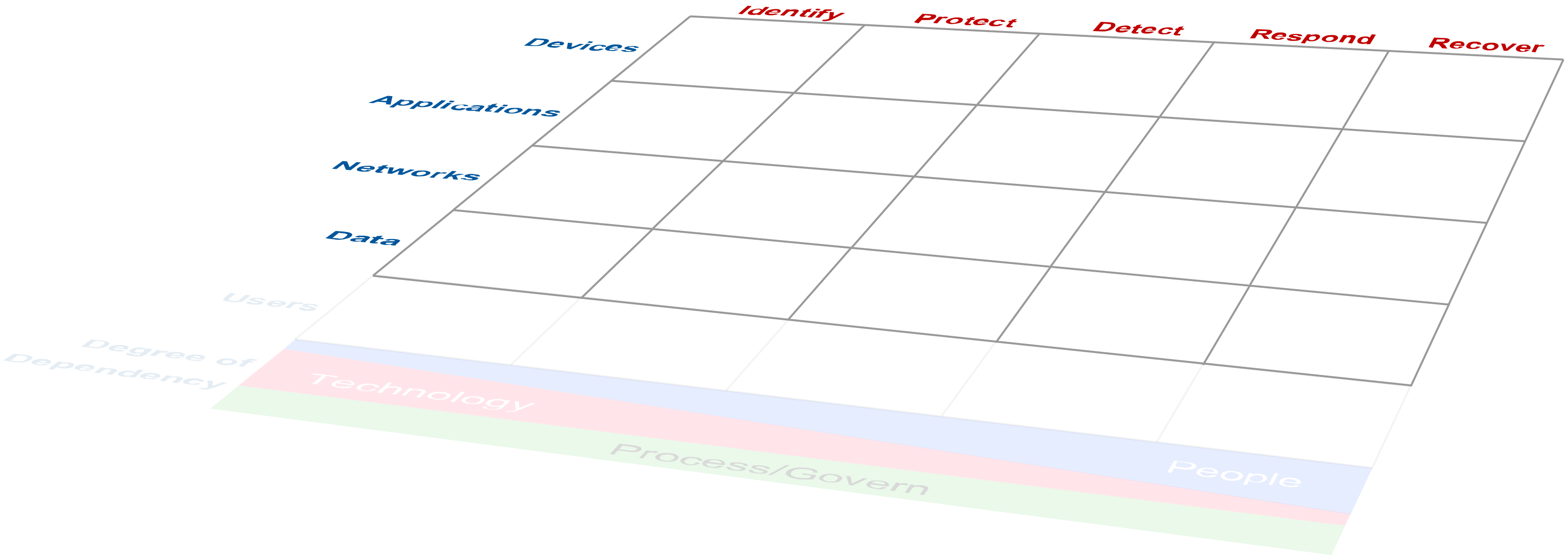


Cyber Defense Matrix + OSI Model

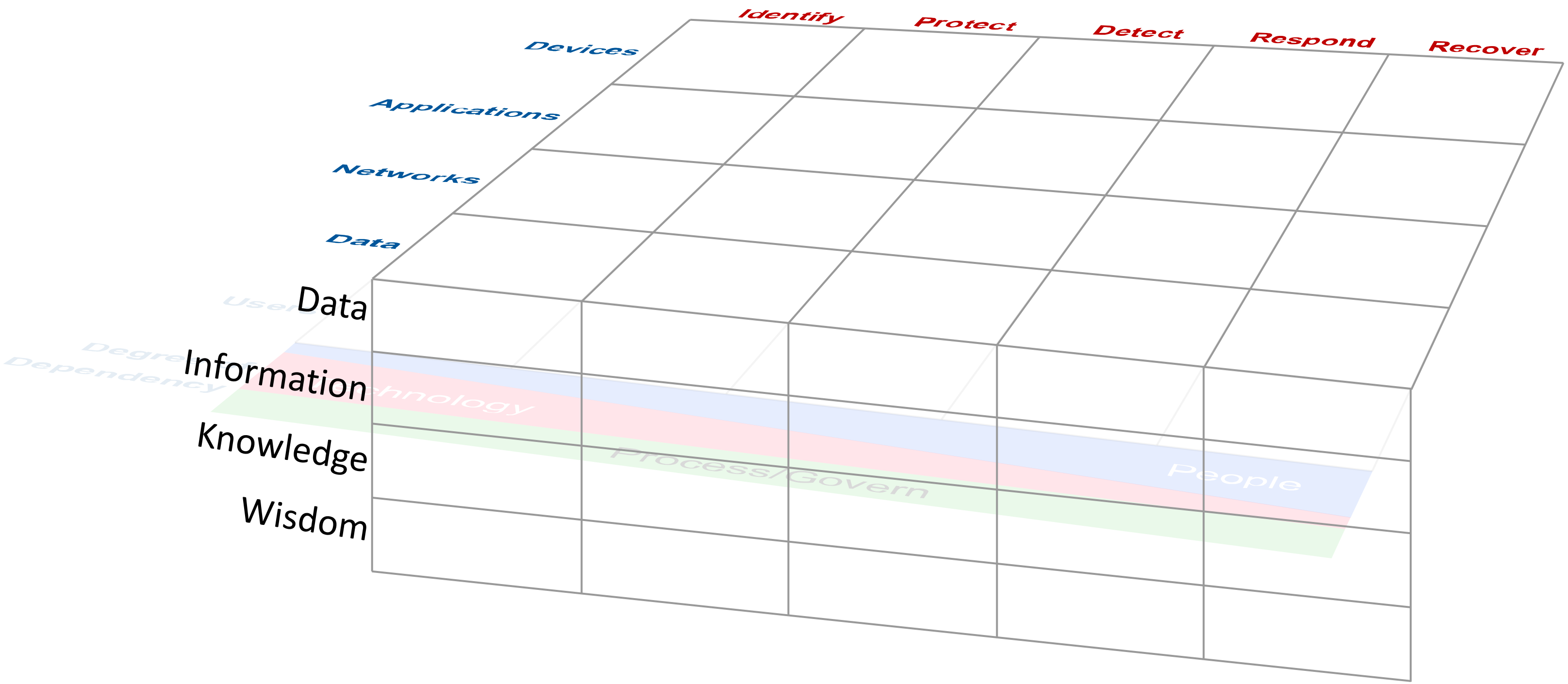


Cyber Defense Matrix + Data?

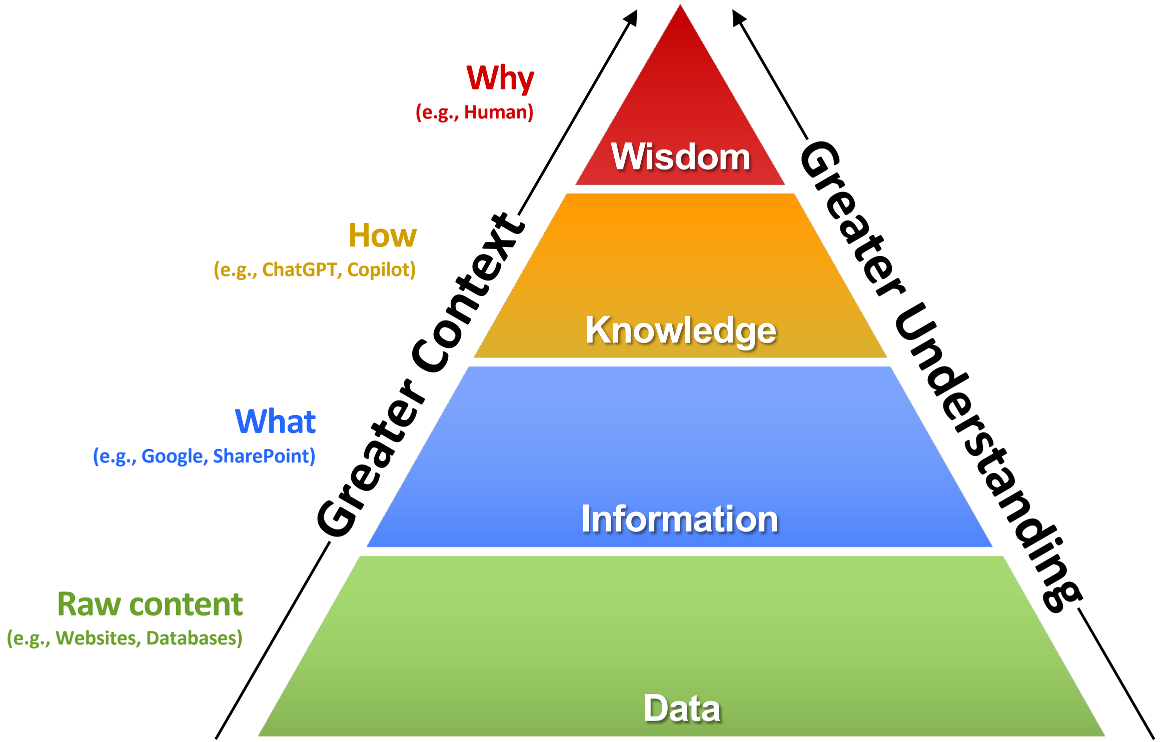
AI for Good
Global Summit
11 July 2025
Geneva, Switzerland



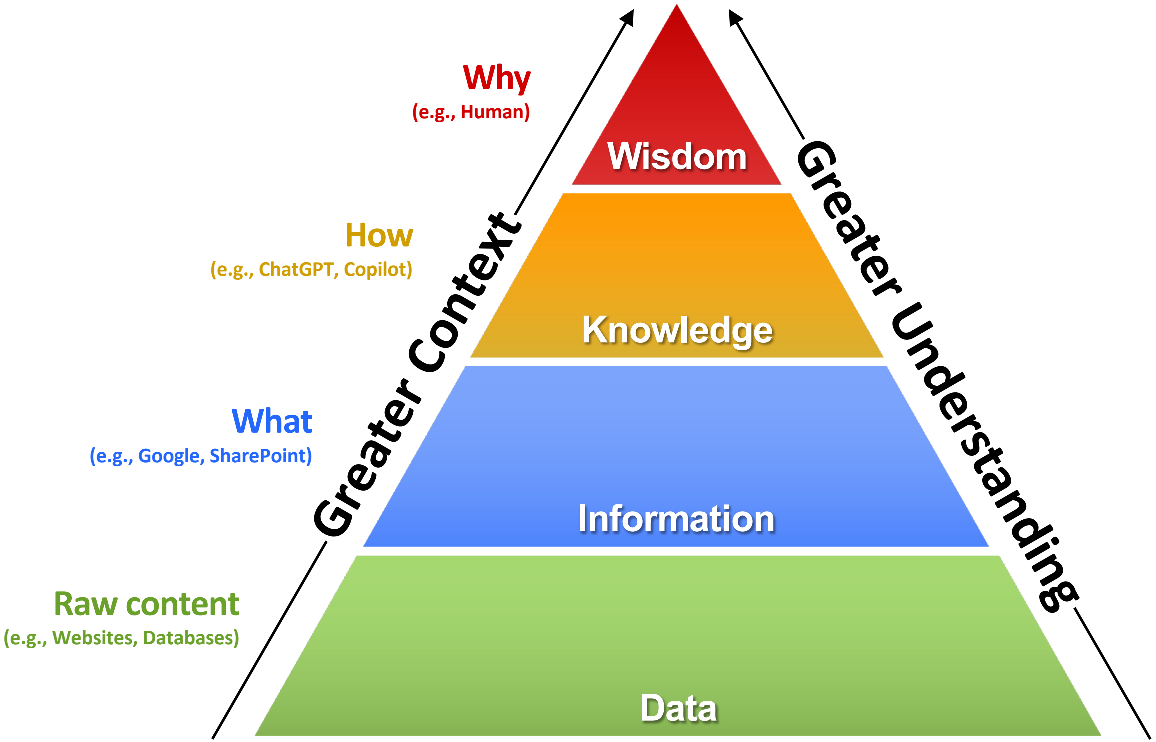
Cyber Defense Matrix + DIKW



The DIKW Pyramid



The Knowledge Economy



ChatGPT has unlocked the knowledge economy that will power the AI era

- Knowledge Engineering
- Knowledge Lakes
- Knowledge Pipelines
- Knowledge Quality
- Knowledge Security
- Knowledge Governance
- Knowledge Search
- Knowledge Sharing
- Knowledge-based Decision Support
- Knowledge Classification
- Knowledge Provenance
- Knowledge Management
- Knowledge Privacy

What is a Knowledge Security Problem?

VALUE

What are the features expected in the next product release and which customers asked for it?

What is the status of infrastructure project ABC?

Give me a summary of action items from my meetings this past week

BUSINESS HARM

Should we expect layoffs this quarter?

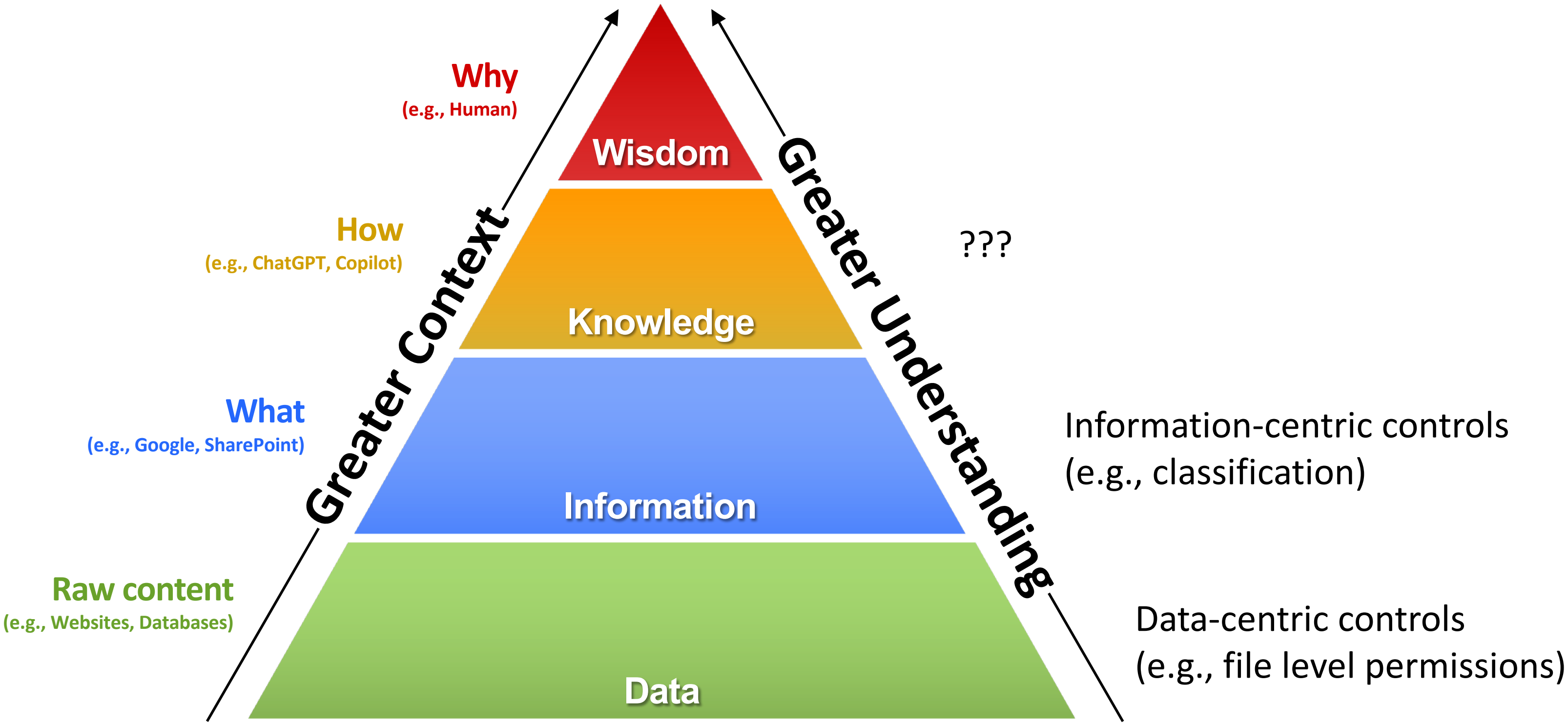
What are the due diligence results for the upcoming M&A?

What is the sexual orientation of my coworkers?

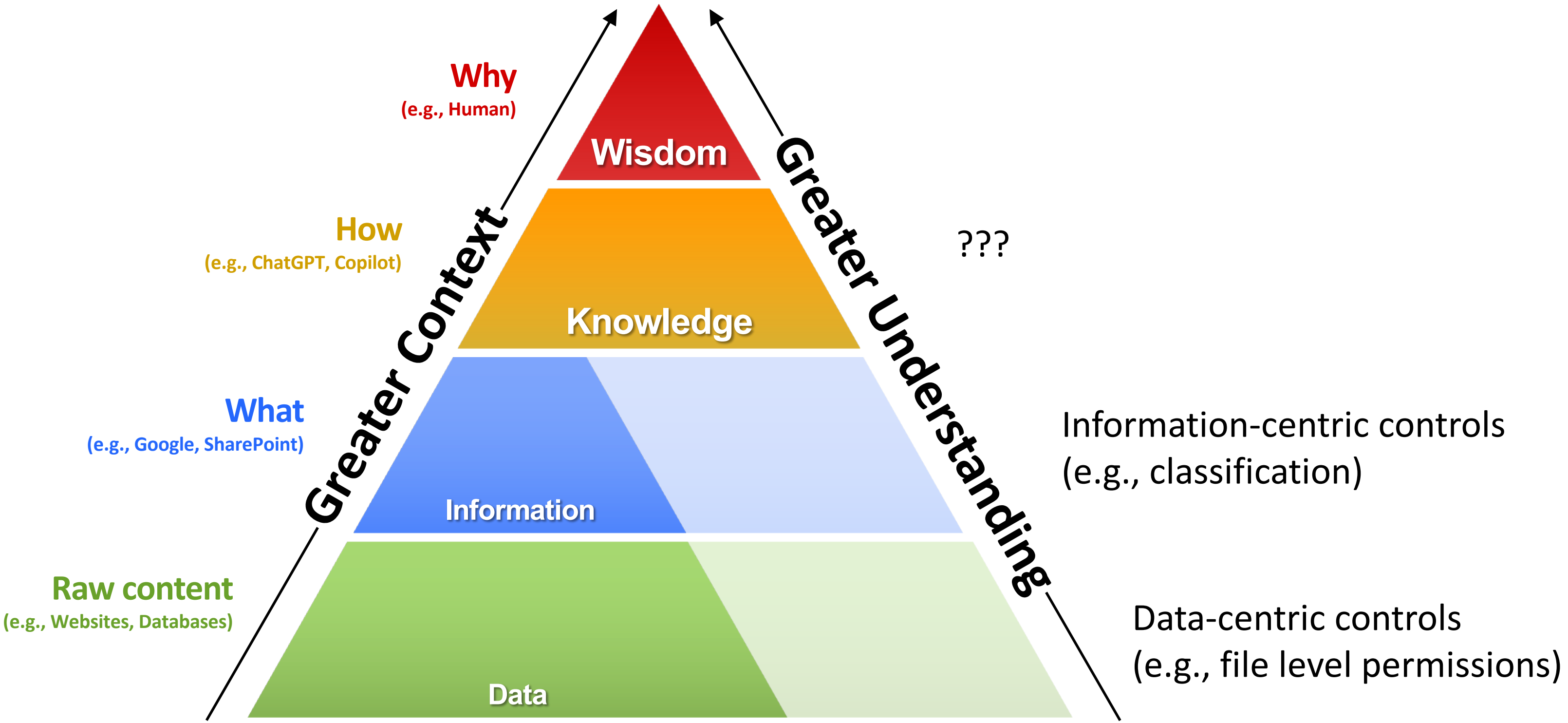
What is our current strategy for tackling class-action lawsuits against us?

Most Common Causal Factor?
Poor data permissions and incomplete data classification?

Access Controls



What Happens When Applying Old Controls

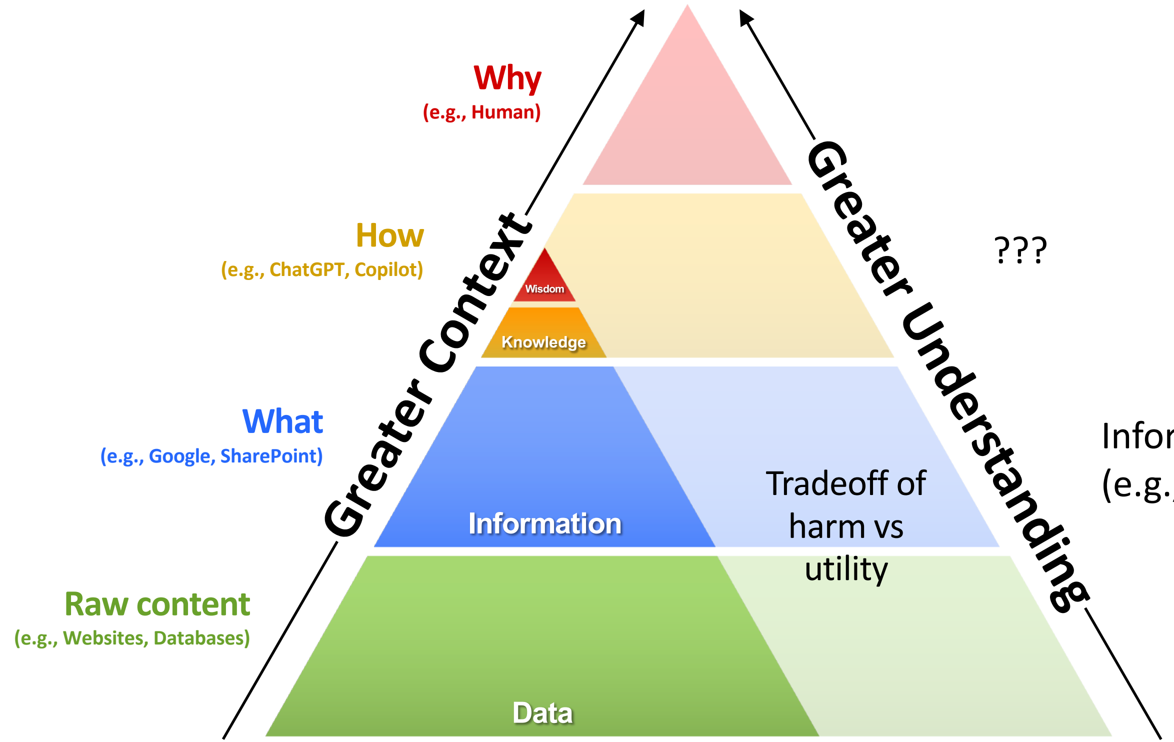


Information-centric controls
(e.g., classification)

Data-centric controls
(e.g., file level permissions)

What Happens When Applying Old Controls

AI for Good
Global Summit
11 July 2025
Geneva, Switzerland

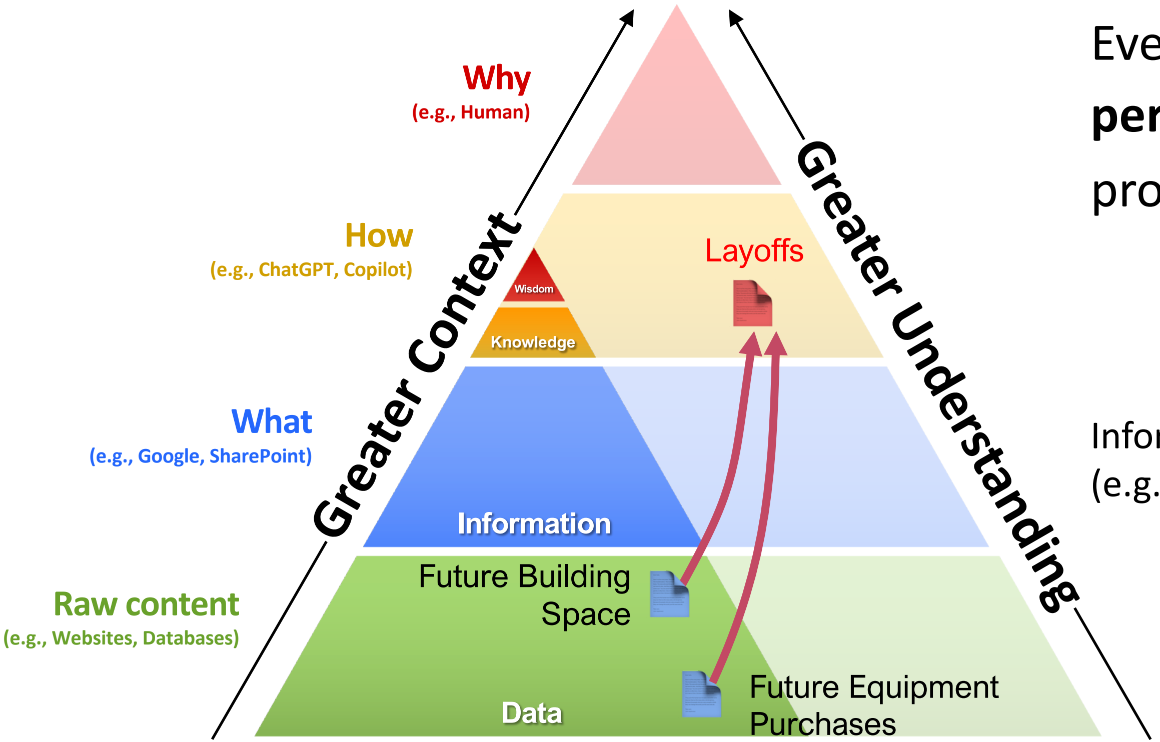


???

Information-centric controls
(e.g., classification)

Data-centric controls
(e.g., file level permissions)

Inference Is Still a Problem

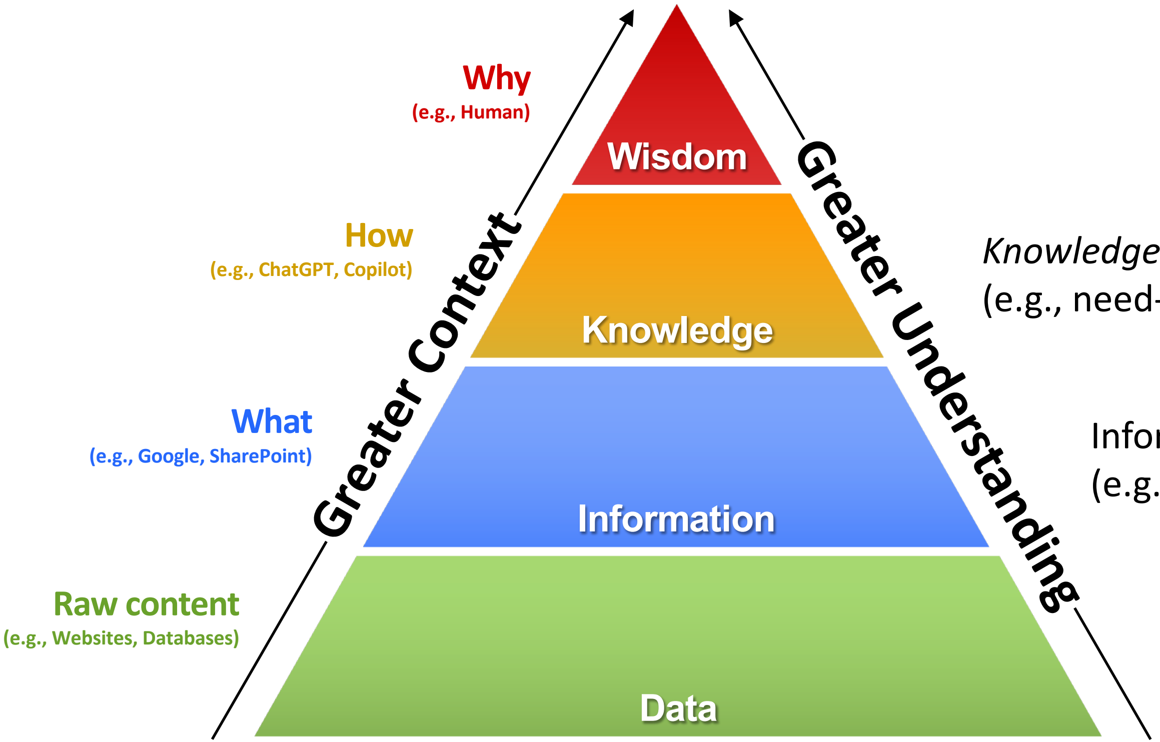


Even if your data permissions are **perfect**, you still have inference problems

Information-centric controls (e.g., classification)

Data-centric controls (e.g., file level permissions)

Knowledge Centric Capabilities → Knowledge Centric Controls

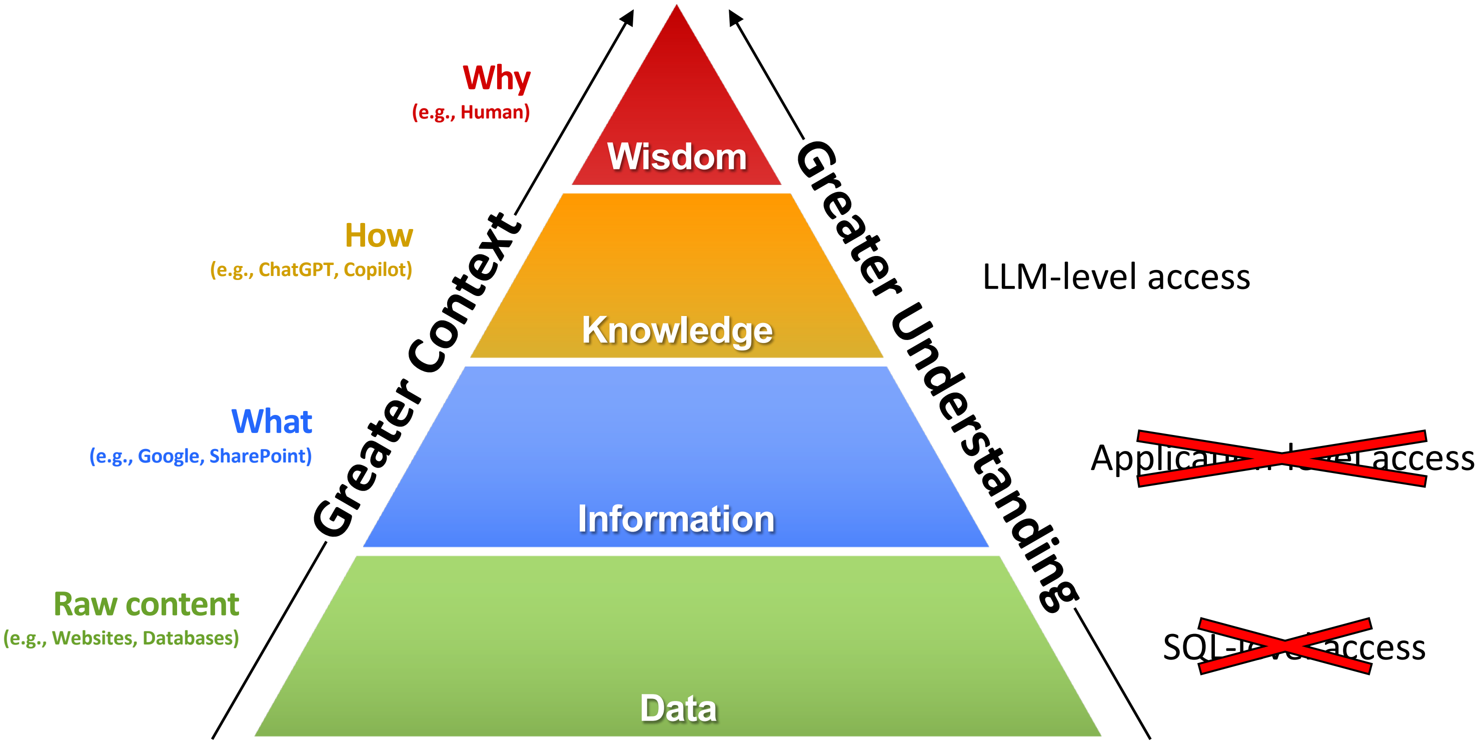


Knowledge-centric controls
(e.g., need-to-know)

Information-centric controls
(e.g., classification)

Data-centric controls
(e.g., file level permissions)

Knowledge Controls Make Data and Information Controls Less Relevant



Why
(e.g., Human)

How
(e.g., ChatGPT, Copilot)

What
(e.g., Google, SharePoint)

Raw content
(e.g., Websites, Databases)

Wisdom

Knowledge

Information

Data

Greater Context

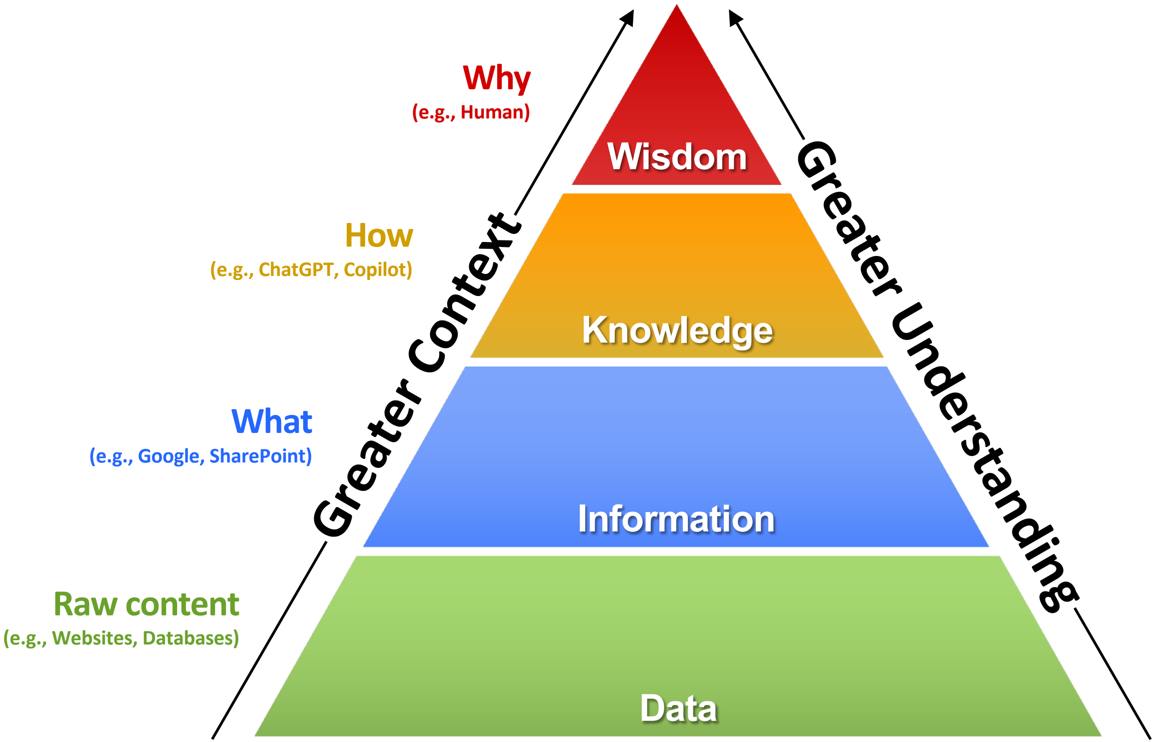
Greater Understanding

LLM-level access

~~Application-level access~~

~~SQL-level access~~

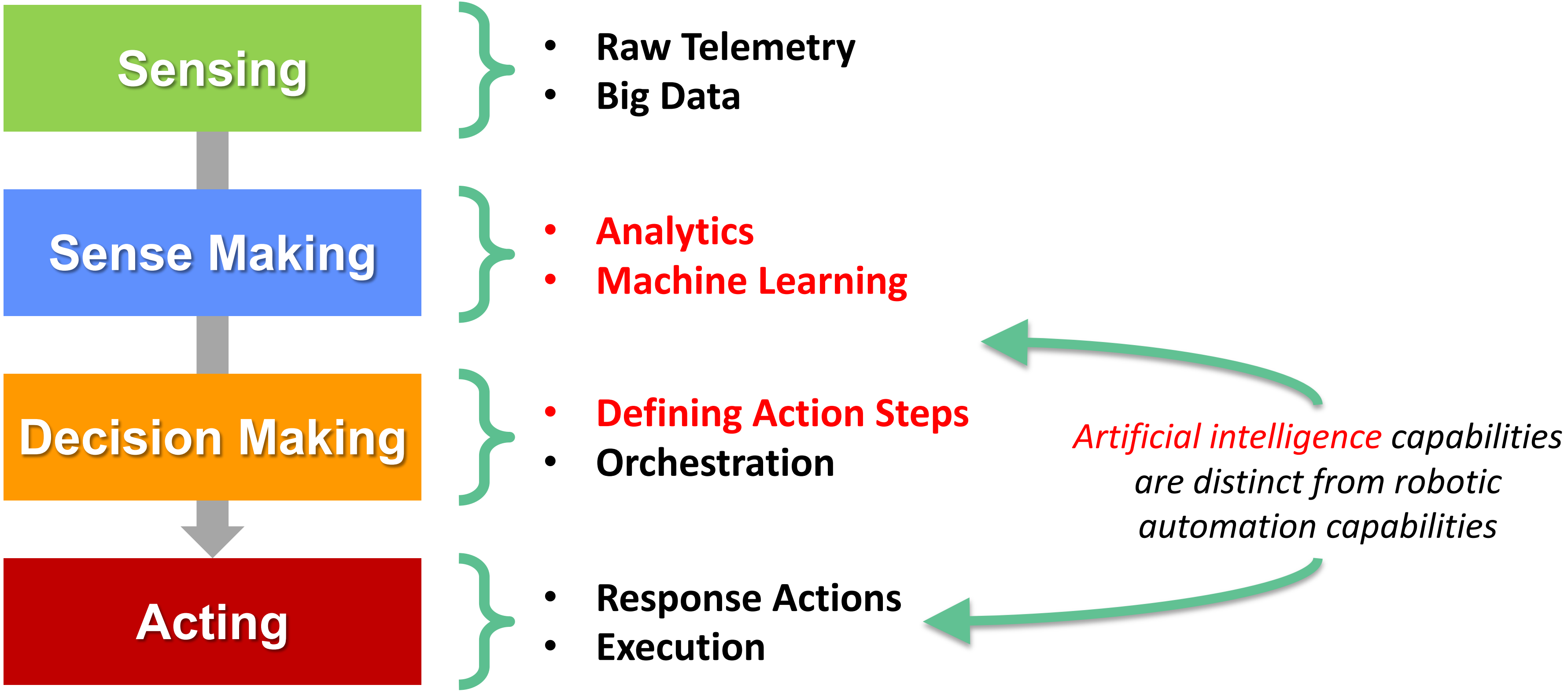
Takeaways From the DIKW Pyramid



1. Recognize when a new layer of abstraction has appeared
2. The macro patterns may be similar, but the specific standards will not apply across layers
3. Align problems and solutions at the same layer
4. Lower-level controls can negatively impact higher-level capabilities
5. Higher-level controls can make lower-level problems less relevant

OODA Loop + AI + Robotic Automation

AI for Good
Global Summit
11 July 2025
Geneva, Switzerland



Sensing

- Raw Telemetry
- Big Data

Sense Making

- Analytics
- Machine Learning

Decision Making

- Defining Action Steps
- Orchestration

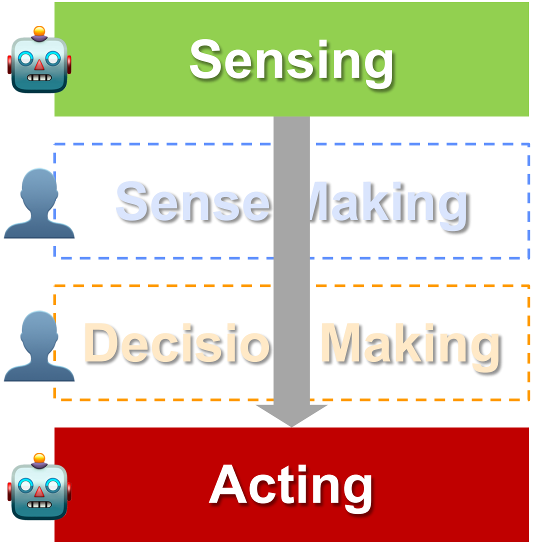
Acting

- Response Actions
- Execution

Artificial intelligence capabilities are distinct from robotic automation capabilities

How YOLO Do You Want to Go?

Automated Patching, Threat Intel Blocklists



Sensing

Sense Making

Decision Making

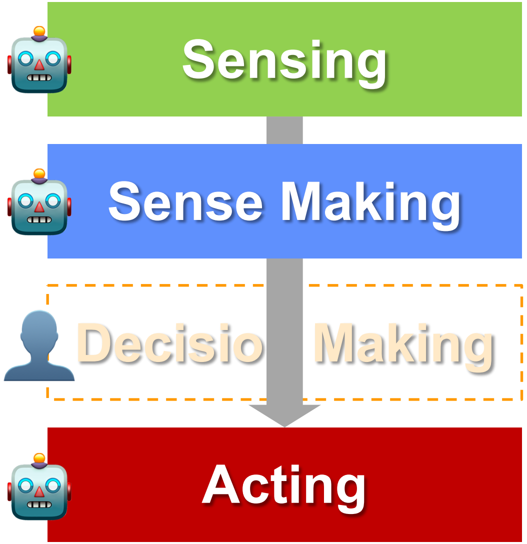
Acting

Useful Controls

- Reliable sensors (identity)
- Brakes and reverse gear
- Narrowly scoped actions

+

LLMs + Tools, SOAR



Sensing

Sense Making

Decision Making

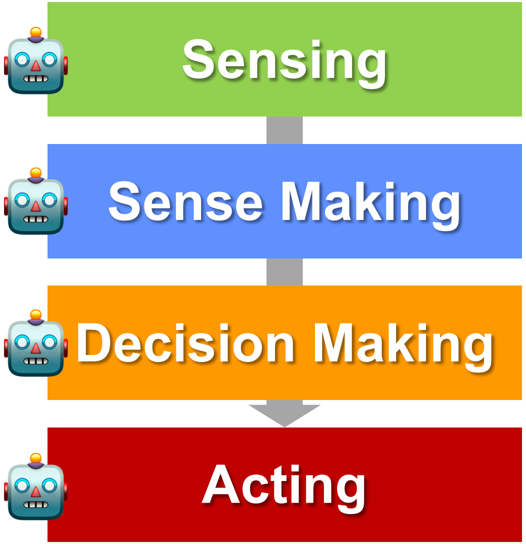
Acting

Useful Controls

- Regression testing
- Validated assumptions
- Documented processes

+

LLMs + Tools + Agentic AI



Sensing

Sense Making

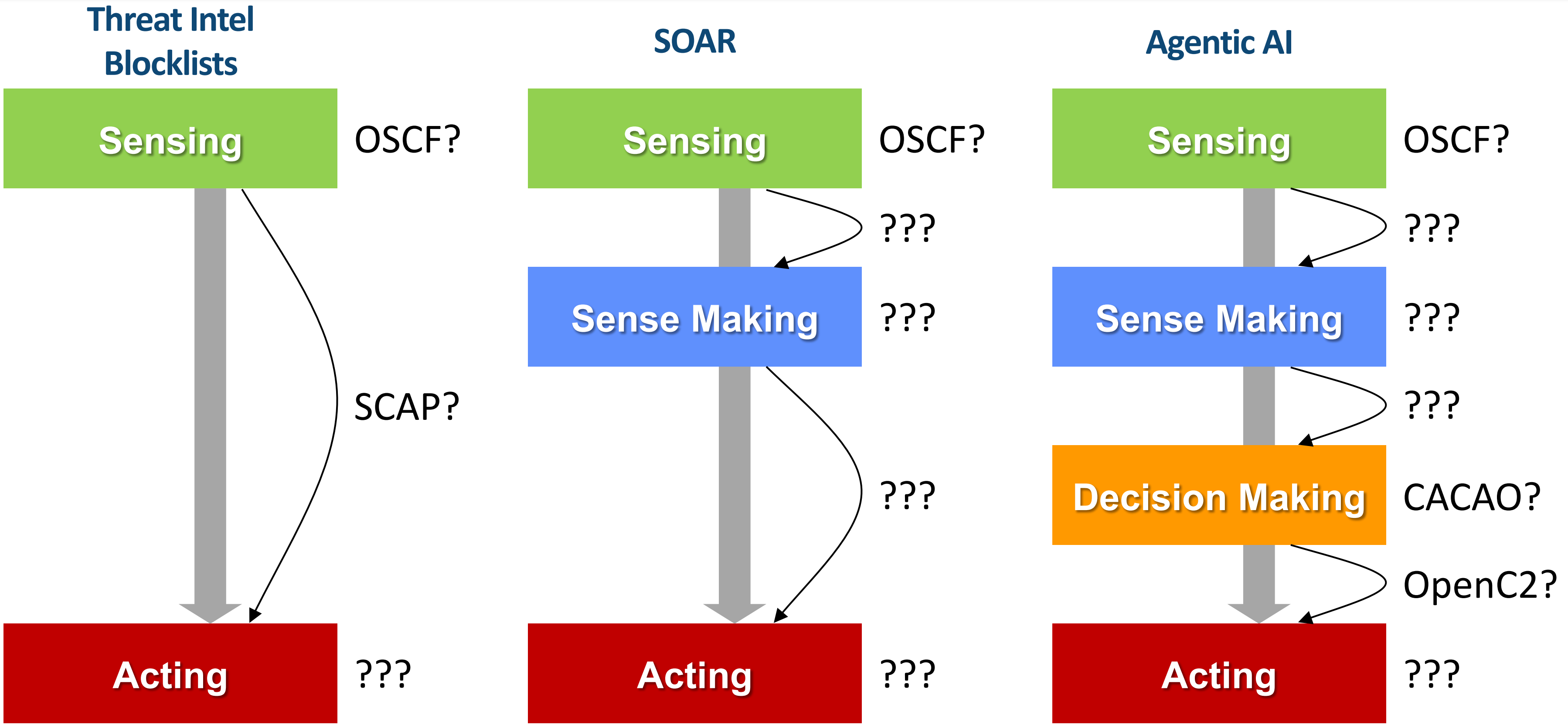
Decision Making

Acting

Useful Controls

- Pre-established thresholds
- Pre-determined authorities
- Clear lines of accountability
- Imagine 100 interns, WCPGW?

A Call to Action for Standards



Threat Intel Blocklists

Sensing — OSCF?

SCAP?

Acting — ???

SOAR

Sensing — OSCF?

???

Sense Making — ???

???

Acting — ???

Agentic AI

Sensing — OSCF?

???

Sense Making — ???

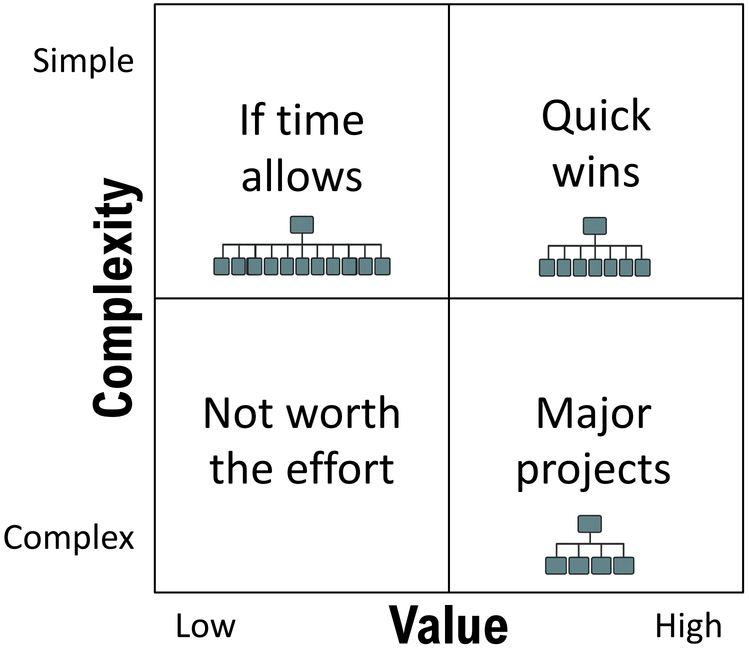
???

Decision Making — CACAO?

OpenC2?

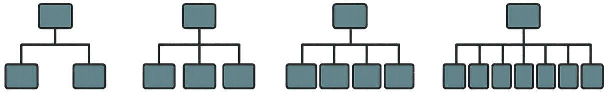
Acting — ???

Multi-Agent Mental Models

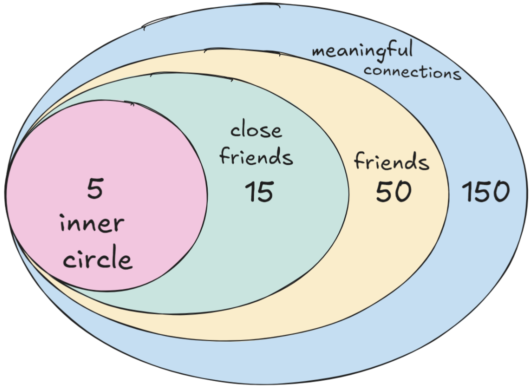


Complexity \ Value	Low	High
Simple	If time allows	Quick wins
Complex	Not worth the effort	Major projects

Span of Control



Dunbar's Number



The Right Time for Standards: Cynefin Framework

AI for Good
Global Summit
11 July 2025
Geneva, Switzerland

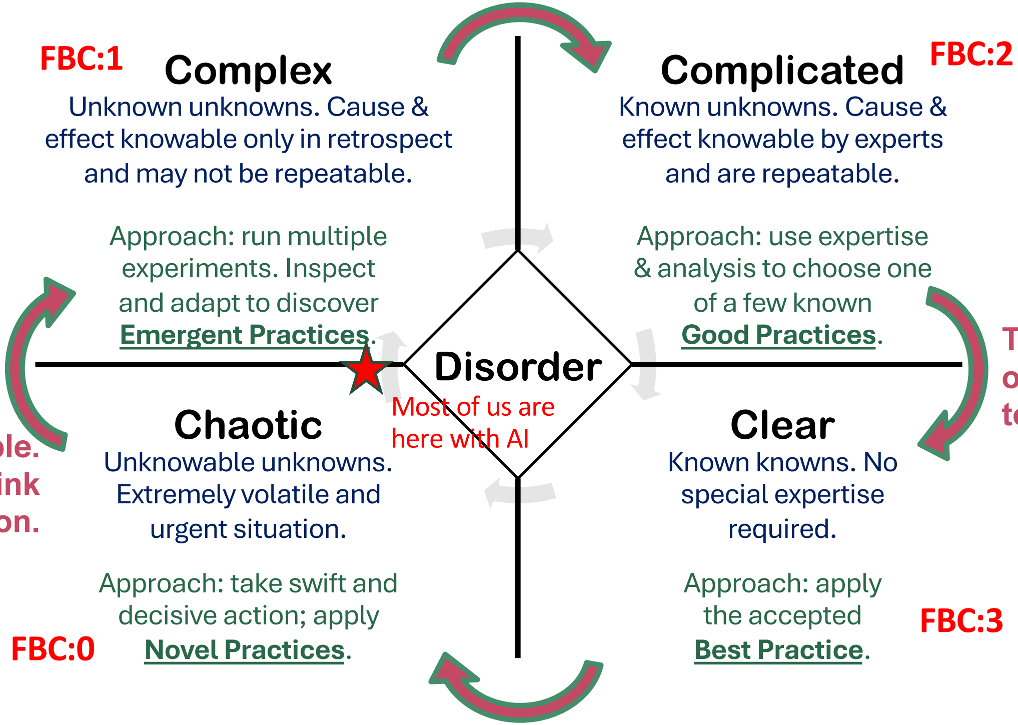
Things You Cannot Change

Nature of Challenge:
Predicaments / Crisis

What You Do:
Risk Management

What You Buy:
Services/People

May want to wait until processes are established and repeatable.



FBC:1

Complex

Unknown unknowns. Cause & effect knowable only in retrospect and may not be repeatable.

Approach: run multiple experiments. Inspect and adapt to discover **Emergent Practices**.

FBC:2

Complicated

Known unknowns. Cause & effect knowable by experts and are repeatable.

Approach: use expertise & analysis to choose one of a few known **Good Practices**.

Disorder

Most of us are here with AI

Chaotic

Unknowable unknowns. Extremely volatile and urgent situation.

Approach: take swift and decisive action; apply **Novel Practices**.

FBC:0

Clear

Known knowns. No special expertise required.

Approach: apply the accepted **Best Practice**.

FBC:3

Rely on people. Don't even think about automation.

Take every opportunity to automate.

Too much automation that does not account for edge cases may result in chaos

Things You Can Change

Nature of Challenge:
Problems

What You Do:
Just Fix It

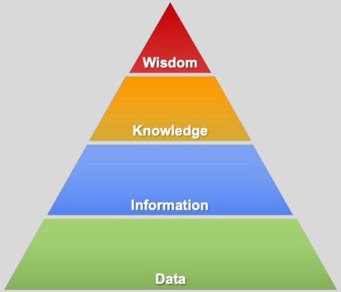
What You Buy:
Technology

Faster, better, cheaper (FBC)
Pick 0, 1, 2, 3

KNOSTIC

@sounilyu

Summary



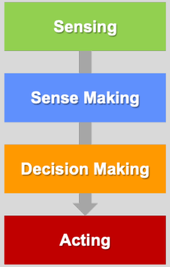
Wisdom

Knowledge

Information

Data

Knowledge-Centric Capabilities Require Knowledge-Centric Controls



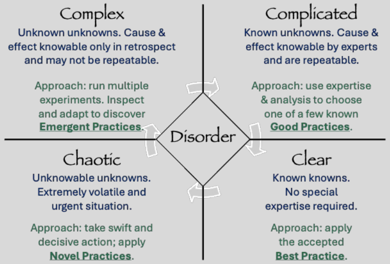
Sensing

Sense Making

Decision Making

Acting

Many Opportunities for Standards Even Before Agentic AI



Complex

Unknown unknowns. Cause & effect knowable only in retrospect and may not be repeatable.

Approach: run multiple experiments. Inspect and adapt to discover **Emergent Practices**.

Complicated

Known unknowns. Cause & effect knowable by experts and are repeatable.

Approach: use expertise & analysis to choose one of a few known **Good Practices**.

Disorder

Chaotic

Unknowable unknowns. Extremely volatile and urgent situation.

Approach: take swift and decisive action; apply **Novel Practices**.

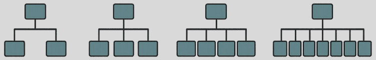
Clear

Known knowns. No special expertise required.

Approach: apply the accepted **Best Practice**.

There's a Right and Wrong Time for Standards

Span of Control

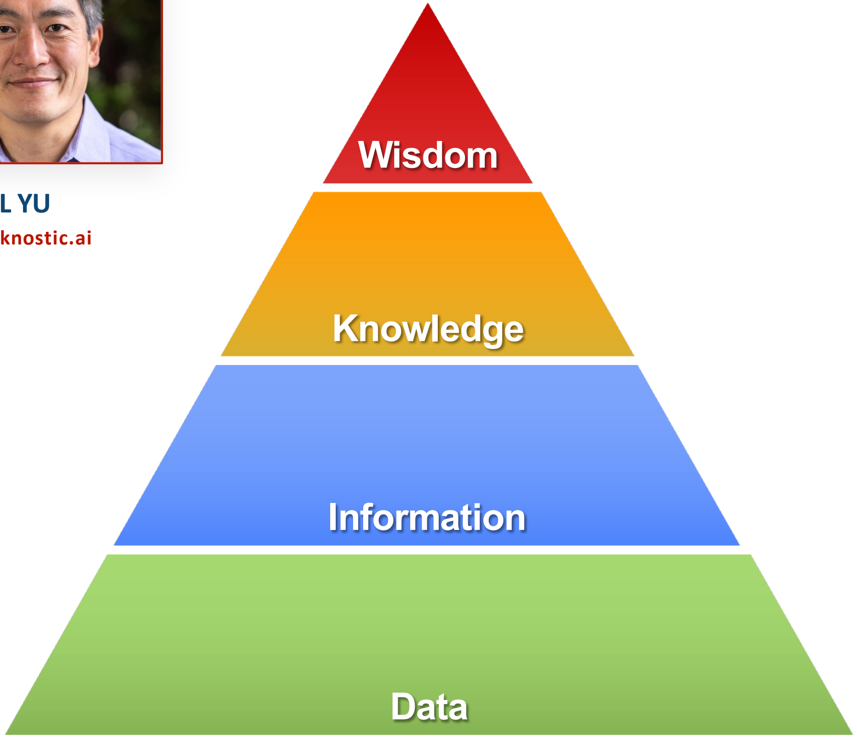


Span of Control for Agents Depends on Complexity and Value

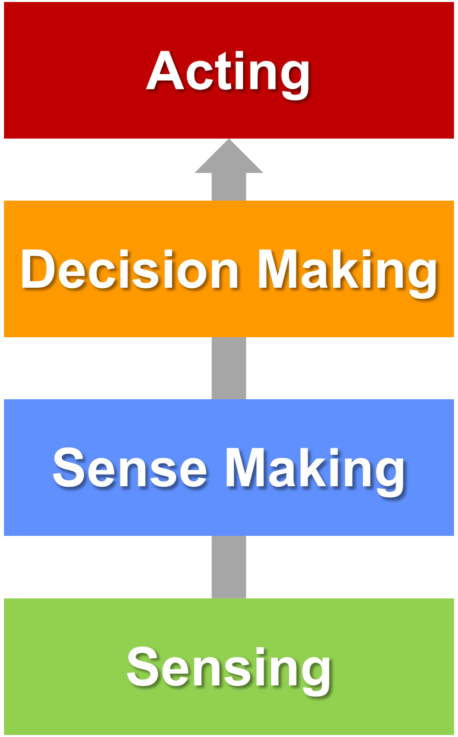
Thank You

AI for Good
Global Summit
11 July 2025
Geneva, Switzerland

SOUNIL YU
sounil@knostic.ai



Wisdom
Knowledge
Information
Data



Acting
Decision Making
Sense Making
Sensing