

HOW TO BUILD TRUSTWORTHY AI

DAVID STRÖM & BJÖRN GENFORS

CADEC 2024.01.18 & 2023.01.24 | CALLISTAENTERPRISE.SE

CALLISTA

TRUSTWORTHY AI

- Trustworthy AI is:
 - Legal
 - Ethical
 - Technically robust

AGENDA

- Ethical AI
 - Requirements
 - Methods
- Technical robustness (briefly)
- Legal AI - EU AI Act
- Summary

THE LETTER



THIS PRESENTATION: EARLY DRAFT

CAN AI BE ETHICAL?

- What is ethical?
- Who makes the decision what is good and bad? Politicians?
- AI is a statistical model based on real world: make the world ethical and AI will follow!

CALLISTA

ETHICS: WHY IT IS BAD FOR BUSINESS

- Academia and politicians love to talk about ethics.
- Ethics is good for some areas, e.g. healthcare and justice
- Free markets adapt to good ethics

IMPACT ON INDUSTRY BY REGULATIONS

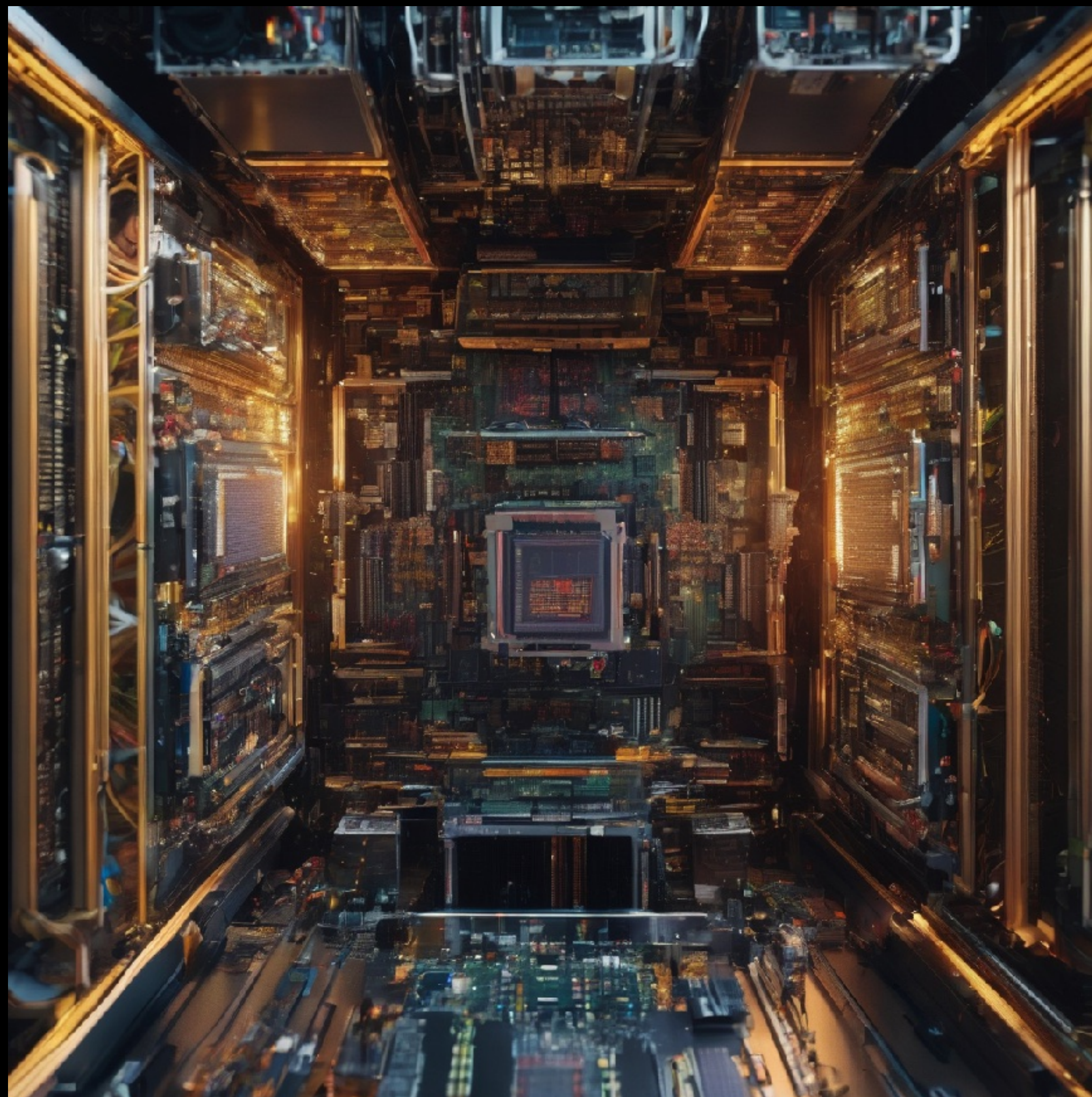
- Well known fact that regulations are bad for business
 - Cost increases for AI businesses estimated to increase by 200% by EU AI Act
-

CALLISTA

| A MESSAGE FROM THE FUTURE



THE SYSTEM



IS THE SYSTEM FREE FROM UNFAIR BIAS?



ACTUALLY, WHAT DATA WENT IN, AND WHAT COMES OUT?



WHO IS IN CONTROL?



DO WE UNDERSTAND WHAT IS HAPPENING?



IS THERE ANYONE ACCOUNTABLE?



WHAT ABOUT EVERYONE ELSE? WHAT ABOUT THE ENVIRONMENT?



ETHICAL AI

ETHICAL FOUNDATION

REQUIREMENTS

PRINCIPLES

FOUNDATION

CALLISTA



ETHICAL REQUIREMENTS

Diversity, non-
discrim. &
fairness

Privacy & data
governance

ETHICAL REQUIREMENTS



Diversity, non-discrim. & fairness

Privacy & data governance

ETHICAL REQUIREMENTS

Human agency & oversight



Diversity, non-discrim. & fairness

Privacy & data governance

ETHICAL REQUIREMENTS

Human agency & oversight

Transparency





Diversity, non-discrim. & fairness

Privacy & data governance

ETHICAL REQUIREMENTS

Accountability

Transparency

Human agency & oversight



Society & environmental wellbeing

Diversity, non-discrim. & fairness

Privacy & data governance

ETHICAL REQUIREMENTS

Accountability

Transparency

Human agency & oversight

HOW TO ACHIEVE ETHICAL AI

GENERAL METHODS

- Diverse teams
- Reach out externally (for more perspectives)
- Foster awareness
- Look to good examples (e.g. healthcare)
- (New) Standards and certifications e.g.
 - Professional codes of ethics
 - Technical standards that confirms the system adheres to safety rules, transparency etc.

TECHNICAL ROBUSTNESS

TECHNICAL ROBUSTNESS (BRIEFLY)

- Guardrails
- Protected against data manipulation
- Fallbacks to rule-based system or human control
- Accuracy
- Reproducibility or predictability



LEGAL AI - EU AI ACT

ETHICAL FOUNDATION


REQUIREMENTS

PRINCIPLES

FOUNDATION

CALLISTA



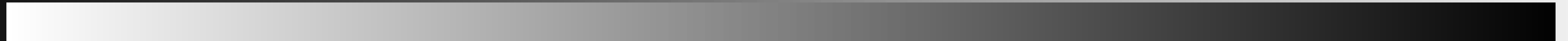


EU AI ACT

WHAT IS AI - LEGALLY?

If-else

Artificial General Intelligence



WHAT IS AI - EU AI ACT DRAFT

- An *AI system* is:
 - all software developed with (at least one of the below):
 - » ML approaches
 - » Logic and/or knowledge based approaches
 - » Statistical approaches
 - which (all of the below):
 - » for human-defined objectives
 - » generates output (content, predictions, recommendations, decisions etc.)
 - » can influence the environment in which it acts



WHAT IS AI? CONCLUSIONS

- Reductio ad absurdum
 - A single if-else clause can be AI
- Some guesswork
 - Internal code not AI. Needs an interface to the outer world.
 - "Influence environment" is a sliding scale. How about a public transport app?
- Take-home message
 - AI Act definition broader than what most people would consider AI

EU AI ACT

- Regulation, not directive
- Still a draft
 - Timeline: expected to come into effect 2025 or 2026
- Brief overview
 - Not complete
 - Fine print exceptions



EU AI ACT

- What?
 - All (non-military) AI systems. Regardless of standalone, or part of a bigger system
 - Foundational models ("GPAI")
- Who?
 - Providers of AI
 - Users of AI and AI generated output
 - Anyone involved in AI value chain (importers, distributors etc.)
- Where?
 - Territorial application as broad as possible





EU AI ACT RISK CATEGORIZATION

EU AI ACT - RISK CATEGORIZATION

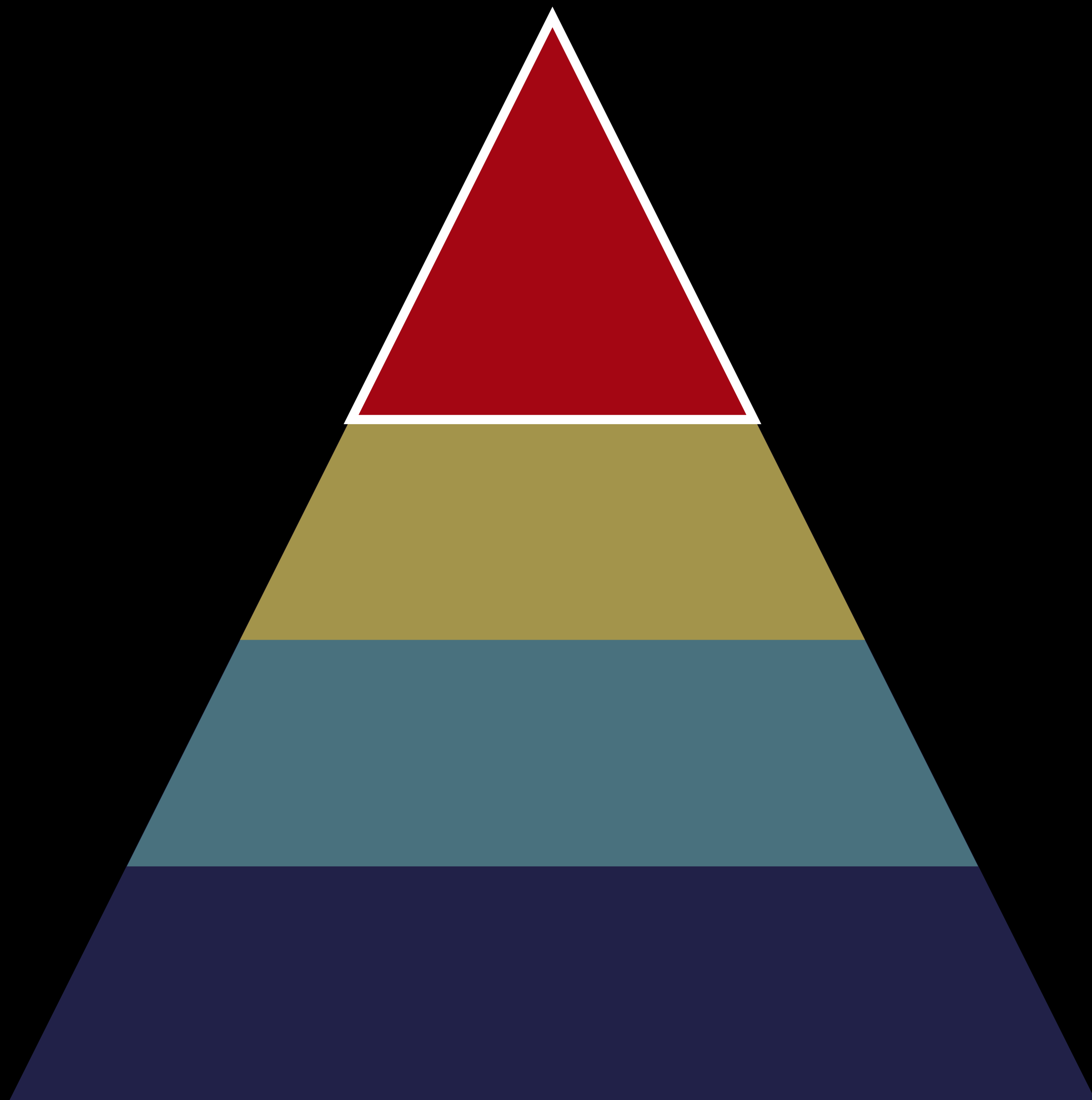


UNACCEPTABLE

HIGH

LIMITED

MINIMAL



UNACCEPTABLE RISK

- Social scoring
- Manipulation of free will
- Wide scope of real-time remote biometric identification
 - (Limited) exception: law enforcement
- ...

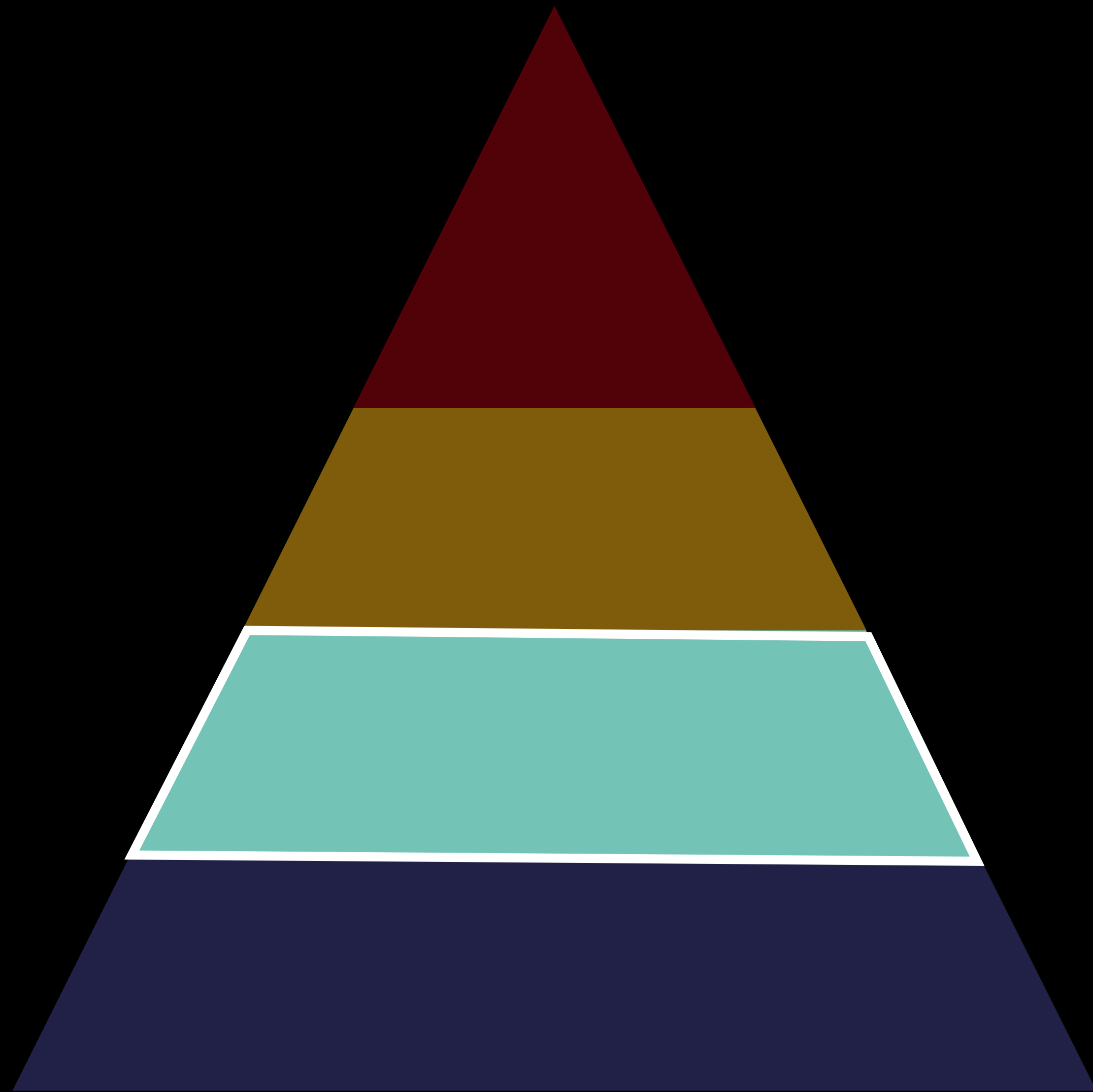
EU AI ACT - CATEGORIZATION



HIGH RISK

- Exercise of governmental authority
- Biometric identification
- Management of critical infrastructure
- Regulated products or safety components thereof
- ...

EU AI ACT - CATEGORIZATION



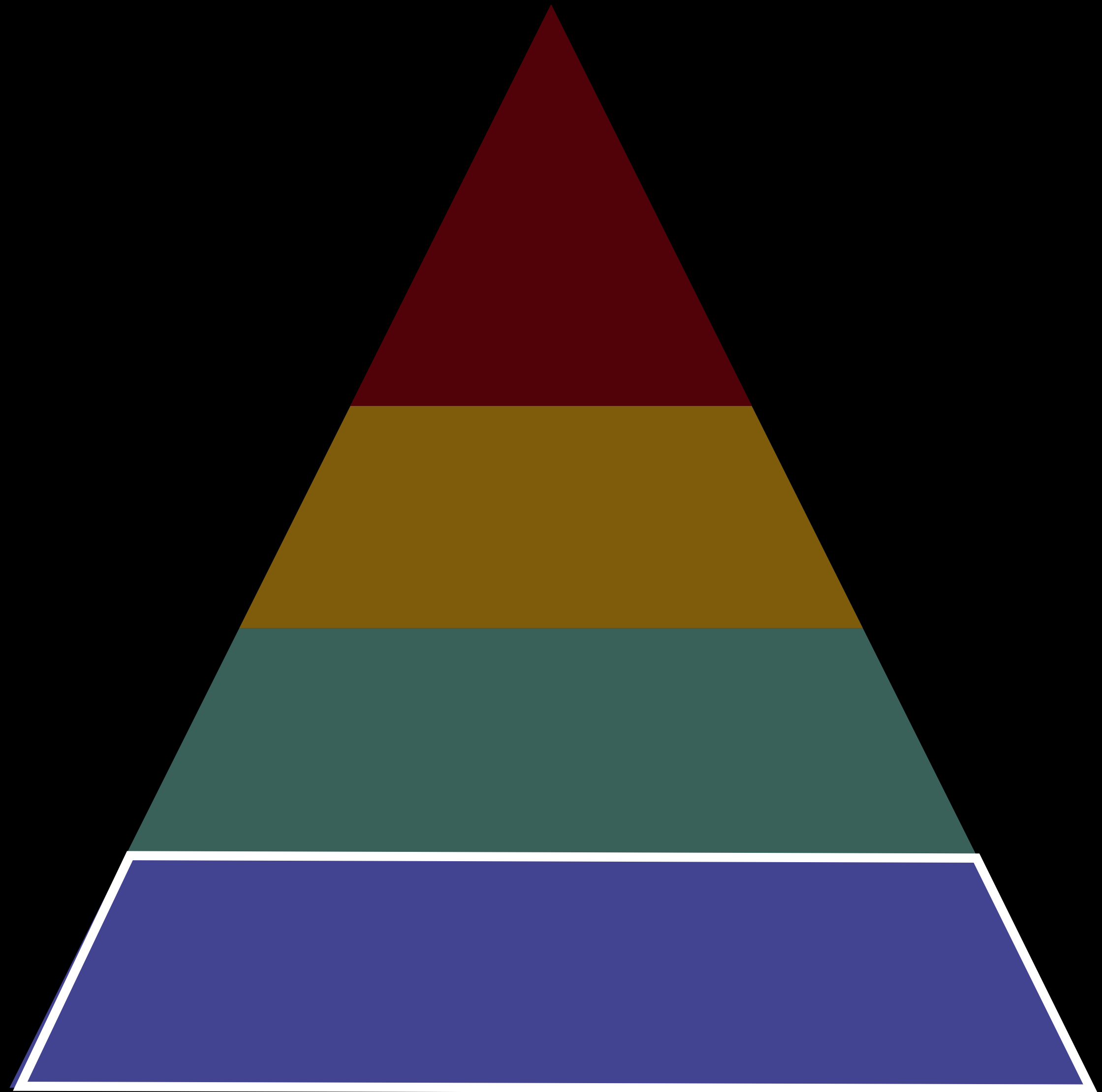
LIMITED RISK

- AI systems that interact with people
- Deep fakes

EU AI ACT - CATEGORIZATION & REGULATION

MINIMAL RISK

- The rest



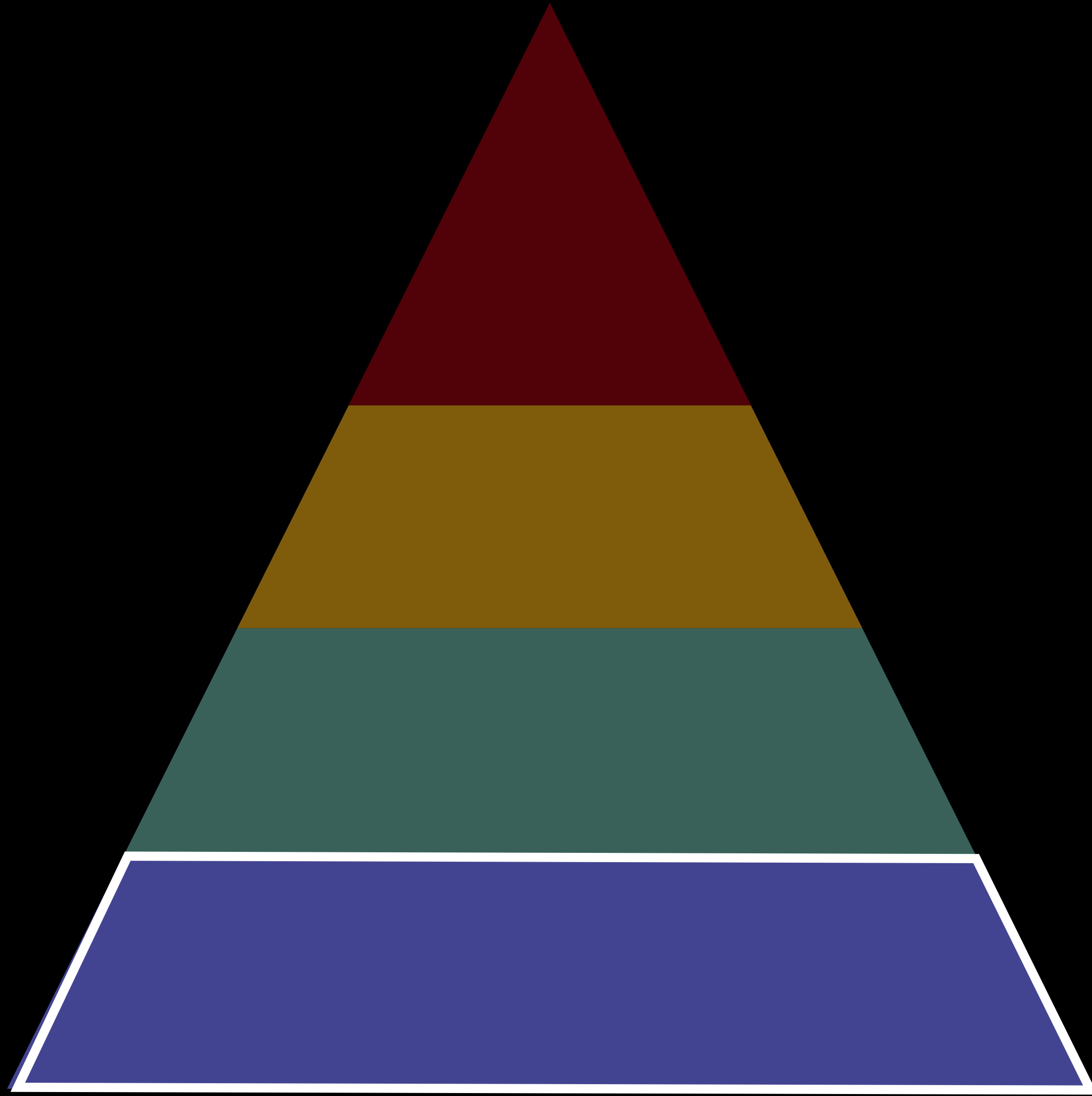


EU AI ACT REGULATION

EU AI ACT - REGULATION

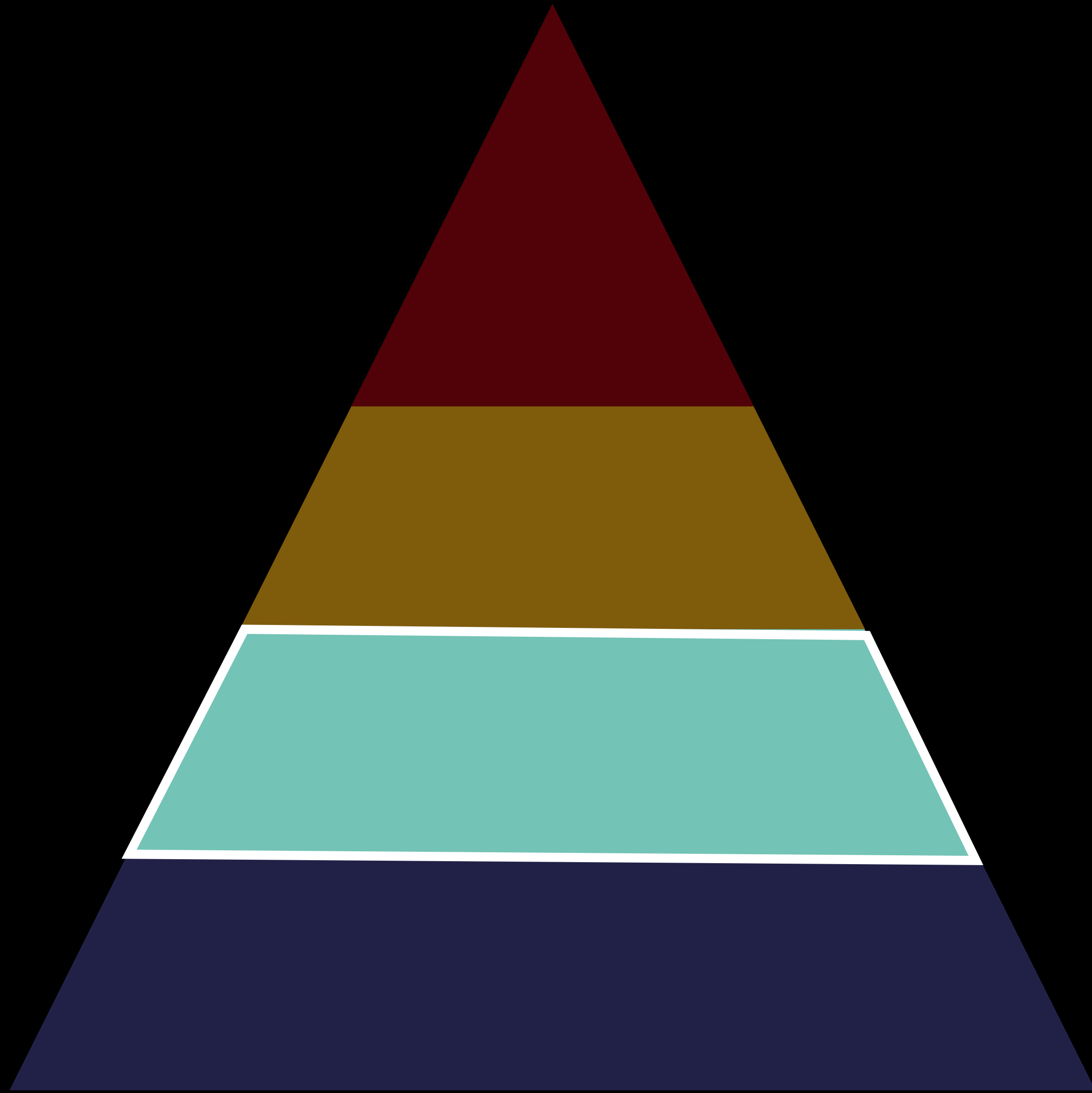
MINIMAL RISK

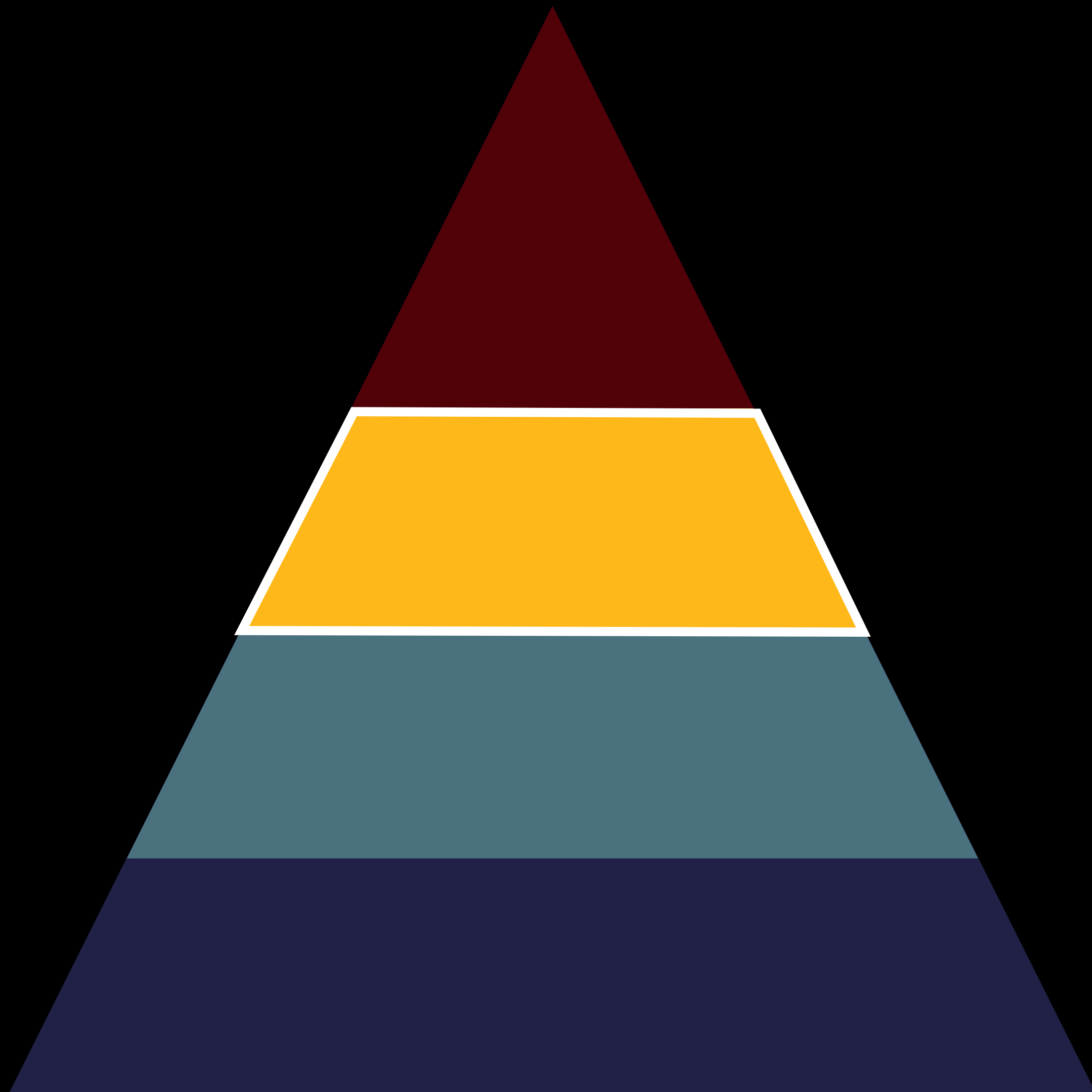
- No regulation



LIMITED RISK

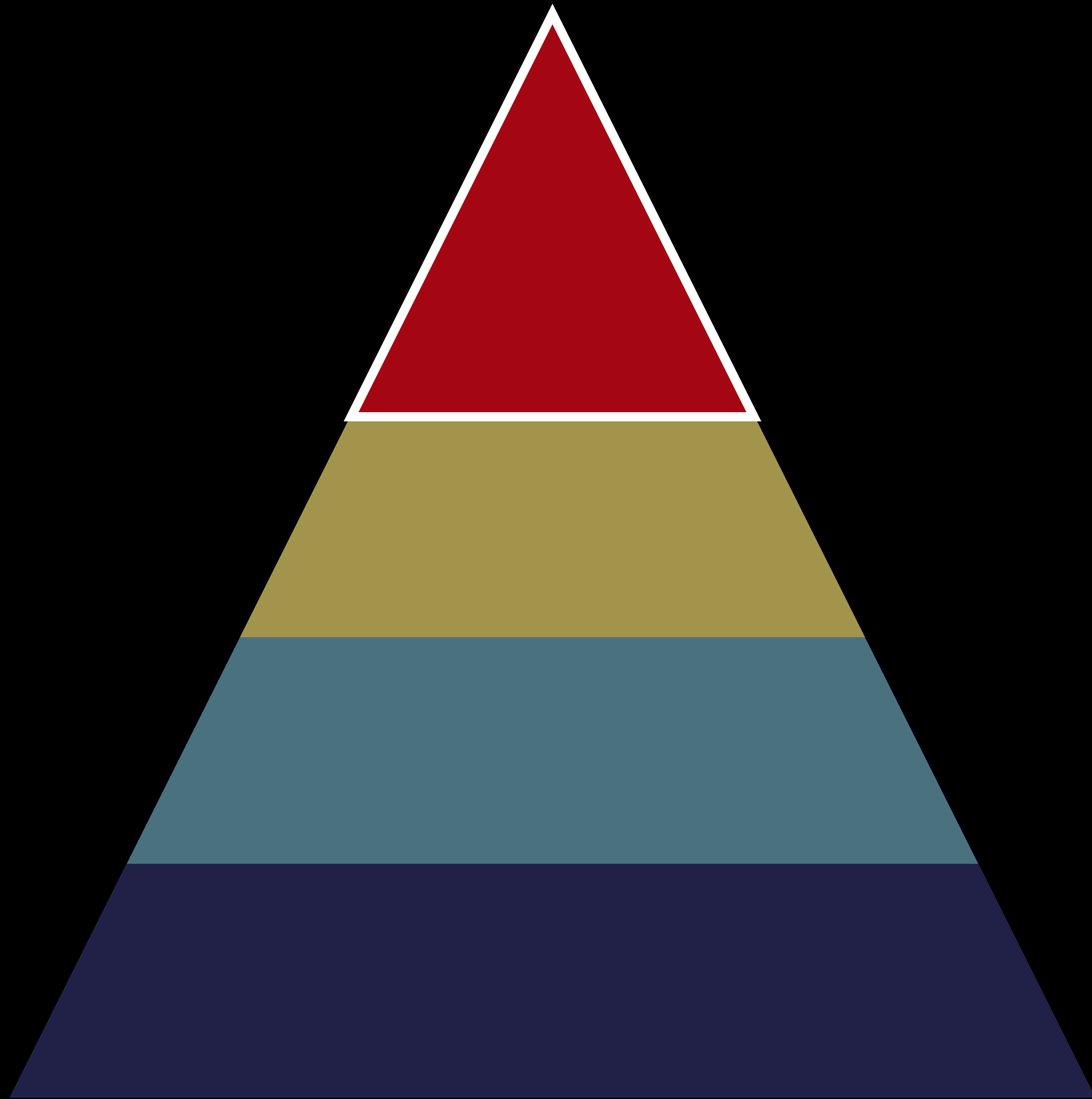
- Transparency





HIGH RISK

- Registration in EU database (standalone AI)
- Hands-on risk management practices
 - Data governance
 - Human oversight
 - Robustness
 - ...
- CE marking
- ...



UNACCEPTABLE RISK

- Prohibited!
- Sanctions up to €35M or 7% of turnover

EU AI ACT - FOUNDATIONAL MODEL REGULATION

- All foundational models:
 - Technical documentation
 - Complying with EU copyright law
 - Disseminating detailed summaries of training data
- High impact foundational models:
 - Additional regulation



EU AI ACT - INNOVATION

- Hard regulation favors big corporations
- Regulatory sandboxing
 - Involves national competent authorities
 - Details TBA



Source: hotpot.ai/art-generator

SUMMARY

| SUMMARY I - WHY DISCUSS TRUSTWORTHY AI?

- The two faces of AI:
 - Endless possibilities
 - Great risks
- AI needs to be trustworthy to gain universal acceptance

| SUMMARY II - CONCLUSIONS

- Trustworthy AI is:
 - Legal
 - Ethical
 - Technically robust
- EU vision is for AI to serve society
 - Broad legal definition of AI
- Short timeline: the time to start adapting to AI Act is yesterday!
 - Let this serve as grounds for making your AI trustworthy

THANKS

USEFUL LINKS

- EU commission AI Act Q&A: https://ec.europa.eu/commission/presscorner/detail/en/qanda_21_1683
- EU Council press release after agreement 8 Dec: <https://www.consilium.europa.eu/en/press/press-releases/2023/12/09/artificial-intelligence-act-council-and-parliament-strike-a-deal-on-the-first-worldwide-rules-for-ai/>
- AI Act: deal on comprehensive rules for trustworthy AI: <https://www.europarl.europa.eu/news/en/press-room/20231206IPR15699/artificial-intelligence-act-deal-on-comprehensive-rules-for-trustworthy-ai>