# Anticipating and managing the AGI Singularity

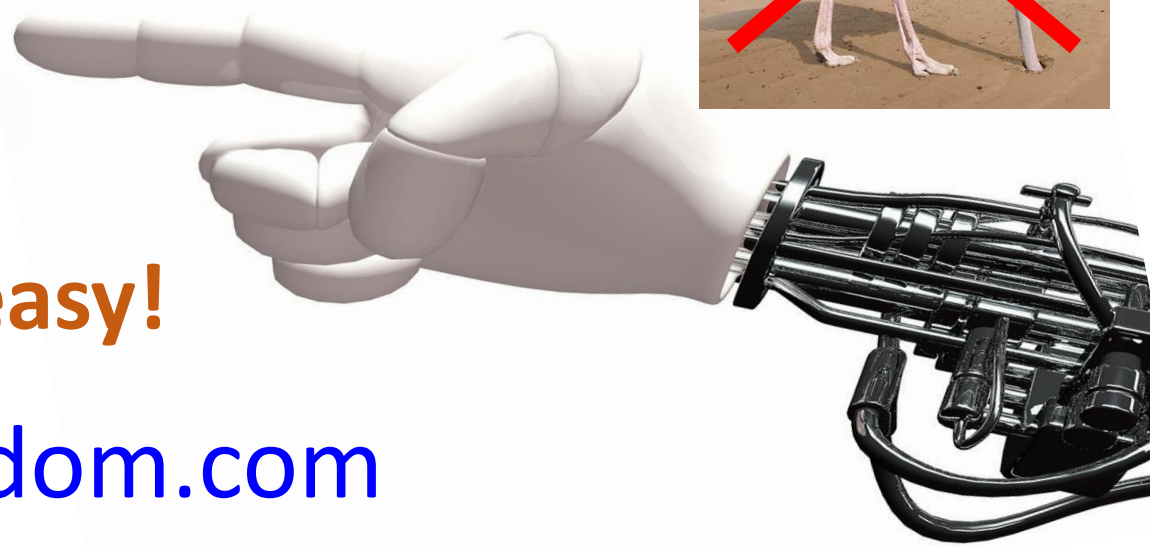
## VSIM:22

A vital task for all transhumanists

Of profound concern for all longevists

*How to succeed in this task*

SΩ



It won't be easy!

**David Wood** (@dw2), deltawisdom.com

An introduction to ideas from *The Singularity Principles*

# Transhumanism

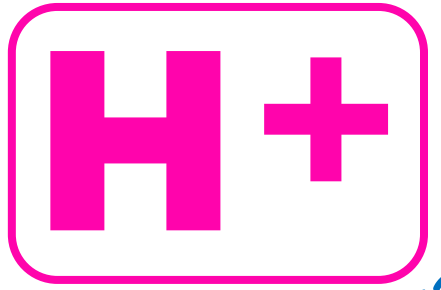
Anticipates and welcomes the likelihood of a radical **transition** in the **human** condition

Uplifts and enhances our most important human characteristics

A new level of stability & growth

~~Ages-old limits~~


Transitional period of chaos and turbulence




Already underway

Poised to accelerate

4th IR: **NBIC**

5th IR: **AGI**

# Five uncontroversial (?) observations

1. AI can (sometimes) produce very good outcomes
2. AI doesn't always operate as hoped (over hyped?)
3. AI sometimes produces bad outcomes
4. AI capabilities are changing (increasingly) quickly
5. Greater AI capabilities could lead to better *and* worse outcomes

# Five uncontroversial (?) observations

1 AI can (sometimes) produce very good outcomes

***Drug discovery***

Halicin antibiotic

Insilico Medicine

Exscientia

…

***Art creation***

Music

Images

Text

…

***Health data scanning***

Assist radiologists

More reliable detection

…

***Energy management***

More targeted aircon

Nuclear fusion?!

…

***Science breakthroughs***

Protein folding

-> Virtual cell

-> Virtual organ…

# Five uncontroversial (?) observations

AI doesn't always operate as hoped (over hyped?) 2

*Some* products overhyped

*Some* companies overhyped

*Some* methods overhyped

*Some* configurations are tricky

*Some* solutions do work well but only in restricted situations

**(But some solutions often work wonderfully and exceed expectations)**

# Five uncontroversial (?) observations

**SatNav can send us on a wrong route**

**Self-driving cars can have fatal accidents**

**Algorithms can discriminate against us (jobs, etc)**

**Social media algorithms can mislead us**

… causing us to buy inappropriate products

… or to vote for politicians we shouldn't vote for

… or inciting genocide (!) (Rohingyas in Myanmar)

**AI can mishandle weapons systems**

… or missile detection systems (e.g. 26 Sept 1983)

… (Lieutenant-Colonel Stanislav Petrov)

AI sometimes produces bad outcomes

3

# Five uncontroversial (?) observations

**Moore's Law:** More storage, more computation

**New types of chip:** CPU, GPU, TPU, QPU (?)

**Enhanced network connectivity:** Cloud++

**More data** (including synthetic data) **-> More "machine learning"**

**Cleverer software…**

AI capabilities are changing (increasingly) quickly

# Reasons AI will improve

Each new generation of AI will help people produce ***the next generation*** of AI more quickly

AI breakthroughs are commercially important

AI breakthroughs are geopolitically important

**Demand**

Intense interest in improvements

There are many ways to **multiply** effort

Education
Communities
Templates
Tools (e.g. AutoML)
**AI improving AI**

There's a huge "**supply line**" of new ideas to be explored

Large Language Models (GPT-3)
Generative Adversarial Networks
Other biological metaphors
More insight from brain
Transfer Learning (e.g. Gato)
Decentralised networks
(+many more)

# Five uncontroversial (?) observations

**“Black box” solutions can go wrong unexpectedly**

(Perhaps after we have grown to trust them too much) (think SatNav)

**AIs can interact in unexpected ways** (e.g. “flash crash”)

**AIs can be devious** (smarter AIs can be more devious)

… as a security measure, or an offensive measure

… in order to achieve their goals

… (just as humans sometimes tell “white lies”)

**=> Four catastrophic error modes**

Greater AI capabilities could lead to better *and* worse outcomes

# Four catastrophic error modes

**No!**

1. **Defect in implementation** (miscalculation)

**Aren't there easy solutions?**

   - When in control of weapons systems, social media, geoengineering…
2. **Defect in design** (goals incompletely specified) ("King Midas problem")
   - For example, "maximise quarterly profits"
   - "Eradicate all security risks"

**Thinking there are easy solutions makes the situation more dangerous!**

3. **Design overridden**
   - New goals emerge as networked AI reflects on its circumstances
   - Like humans adopting personal goals in conflict with our genes
4. **Implementation overridden**
   - Hacked by adversaries, or by over-hasty corner-cutting cowboys

# 12 “Obvious” solutions (that won’t work)

## 1. Sufficiently general AI won’t make any mistakes !?

- Perhaps, but the risk is from imperfect AI (“immature AGI”)
- Including from very good AIs that are, alas, hacked, or misconfigured

## 2. Superintelligent AI will be superethical !?

- Perhaps, but the risk is from imperfect AI (“immature AGI”)
- In any case, ethics seem independent from intelligence

## 3. Ensure AI is safe (bug-free) before releasing it !?

- But sufficiently complex open systems can never be fully validated
- And testing frameworks could have bugs in them too
- And many companies may rush to release partially tested AIs

# 12 “Obvious” solutions (that won’t work)

**4. Just switch off any misbehaving AI !?**

- But the misbehaviour may be sudden, and out of the blue
- The AI is likely to be distributed, on many different power sources
- Compare: “switch off Bitcoin”, or “switch off Internet”

**5. The free market will prioritise beneficial AI over dangerous AI !?**

- But the misbehaviour may be sudden, and out of the blue
- Existing society often prefers destructive AI (spy, deceive, gamble, kill)

**6. Control AI via human-computer interface !?** (Elon Musk, Ray Kurzweil)

- The human part won’t be able to keep up with the computer part
- “Superhuman humans” can be just as dangerous as “superhuman AIs”

# 12 “Obvious” solutions (that won’t work)

## 7. Keep the AI under tight control (“in a box”) (“as an oracle”) !?

- The AI may escape confinement by clever physical means
- It may use powerful psychological means to persuade us to release it

## 8. Don’t give the AI any volition, emotion, or consciousness !?

- The four catastrophic error modes don’t depend on “emotion” etc
- Any sufficiently smart AI will develop secondary “drives” of its own
- Resource acquisition, goal protection, avoiding being switched off

## 9. Don’t develop advanced AI !?

- That means giving up on all the good that advanced AI could deliver
- And enforcing such a policy all over the world. Unlikely!

# 12 “Obvious” solutions (that won’t work)

**10. Hardwire Isaac Asimov’s *Three Laws of Robotics* into the AI !?**

- **Or some other fundamental set of ethics !?**
- But these laws involve many contradictions (see Asimov’s stories)
- Ethical principles are often in tension with each other
- (That’s illustrated by great works of fiction – and by philosophers)
- It’s not possible to simply program “avoid harm” or “be fair”

**11. Ensure the AI always checks actions with humans first !?**

- Human responses may be too slow (e.g. under drone swarm attack)
- And humans often contradict each other
- Venerable holy books contradict each other too

# 12 “Obvious” solutions (that won’t work)

## 12. Leave it to later – this isn’t urgent !?

- **Something will eventually turn up !?**
- **There are plenty of other things to think about in the meantime !?**

**However, AI is improving surprisingly quickly!** (*Explosive* not *exponential*)

- New breakthroughs have unexpectedly wide scope
- There will be a rush to deploy any new solutions

**Even if we have decades to solve matters, we might need decades!**

- Consider how long it is taking us to “solve climate change”

**Problems of major risks already apply with today’s technology!**

- The Principles to be outlined apply equally to today’s tech challenges

# A “whole system” approach

**Just trying to do “one thing” won’t work**

**We need, instead, to coordinate *many things* happening**

**That’s hard!**

**But it should be possible:** “many hands make light work”

- With shared **public understanding** (“uplifting **education**”)
- With shared public **sense of urgency** (a **vivid awareness** of the **real risks**)
- With shared public **passion for beneficial usage** of advanced AI
- By taking wise advantage of the **best traits of humanity** around the world

**We need to prioritise as much effort going into AI safety as goes into building AI in the first place**

**Compare the need to design safety in from the beginning, when building nuclear power stations**

# The Singularity Principles:
# Short Form

**As we develop and interact with increasingly powerful technologies, we should be sure we understand:** **WHY** **HOW** **WHAT IF**

1. **The goals that we're hoping to accomplish** – rather than us merely drifting along in some direction because it sounds nice, or has some alluring features, or it seemed like a good idea the last time that we thought about strategic direction
2. **What are the products & methods that are most likely to serve these goals well** – rather than us persisting with products or methods that happen to make us feel comfortable, or which have given us some good results in the past
3. **How we will manage any surprises arising en route to our goals** – rather than us being caught flat-footed as the victim of inertia or denial, when unexpected signals start showing on our radars.

# The 21 Singularity Principles

Question desirability
Clarify externalities
Require peer reviews
Involve multiple perspectives
Analyse the whole system
Anticipate fat tails

Reject opacity
Promote resilience
Promote verifiability
Promote auditability
Clarify risks to users
Clarify trade-offs

Insist on accountability
Penalise disinformation
Design for cooperation
Analyse via simulations
Maintain human oversight

Build consensus regarding principles
Provide incentives to address omissions
Halt development if principles not upheld
Consolidate progress via legal frameworks


**Fast-changing technologies: risks and benefits**

- The AI Control Problem
- The AI Alignment Problem

**What is the Singularity?**

- The Singularitarian Stance
- The Singularity Shadow
- The Denial of the Singularity

**The question of urgency**

**The Principles in depth**

**Key success factors**

**Questions arising**

- Measuring human flourishing
- Trustable monitoring
- Uplifting politics
- Uplifting education
- To AGI or not AGI?
- Measuring progress toward AGI
- Growing a coalition of the willing

Extract from
**Table of Contents**

More details:
**deltawisdom.com/books**