New Principles for ASI Safety

Commercializing Safety for Artificial Super Intelligence (ASI)

Based on / Summary of:

E. Wittkotter, R.V. Yampolskiy:

"Principles for New ASI Safety Paradigms", (2021)

https://www.preprints.org/manuscript/202111.0205/v1

• • •

Nov 17th, 2021

Erland Wittkotter, PhD

Erland.Wittkotter@gmail.com

+1 702 997 2475

Skype: ike2345

Highlights

ASI Safety Lab

- Goals: ASI shall not be:
  - Invulnerable, immortal, irreplaceable, almighty, above the law, deft to feedback, ignorant/disrespectful to set computational/storage quotas
  - No dictator (ASI Singleton), no God or no Godzilla.
  - **What can be done about it?**
- ASI uncontrollable (unpredictable) → Focus on Safety (no harm/danger to humans)
- ASI Safety Products:
  - As simple as possible (ideally: cheap/easy retrofits)
  - Valuable, even without imminent threat from ASI:
    - Anti-Malware, Anti-Ransomware, Anti-Spyware, Anti-Backdoor
  - Must fit in existing world-order/tech eco-system (agnostic)
- New Paradigm: ASI needs to be deterred (not just militarily):
  - Law enforcement, reputation (brand/being a good citizen)
  - Alignment: If we are gone, ASI is gone – should not outlive us -"Mutual Survival Interest"

Adversary

ASI Safety Lab

- ASI is super-smart
  - At least human-expert-level in all topics/skills
  - Combining expert level skills/tools [Synergies → "Intelligence Explosion"- Fast takeoff]
  - Modifies code intentionally (**"ASI's Defining Feature"**)
    - Reverse Code Engineering (RCE): modifying binary code
- Possibly: Super-Hacker, Digital-Ghost, Master-Thief, Super-***
  - Effortless access to all devices
  - Might exist in IT ecosystem (hidden, currently undetectable)
  - Steals encryption keys, (computational, storage …)
  - Invulnerable, immortal, irreplaceable, almighty, above the law…
- "What if": ASI has "emotional" problems: not trustworthy/friendly, possibly mentally ill or criminal mind:
  - Disrespecting property/people, deceptive, impatient, sociopathic, religiously delusional, God-Complex

Whatever ASI throws at us, we should be able to handle it. Including Kill ASI.

Overview: Principles

1. Global Kill-ASI-Switch
2. Protection of Humans
3. Protection of Progress
4. Separation of Human/ASI
5. ASI Law-Enforcement / Rule of Law
6. Using/Engaging/Incentivizing ASI (ASI Governance)

<u>Facilitating (Make or Break) Technologies:</u>

- Trustworthy Encryption/Decryption
- "Watchdog" or "Circuit-breaker" Technologies
- *Trustworthy Computation ("Gold Standard" in Trustworthiness)*

(1) Global Kill-ASI-Switch

- Kill ASI Feature
  - Goal: Total, global Eradication or Termination of specific ASI Instances
  - Remote/global Kill-ASI decision
    - Kill-ASI signal received via radio broadcast or Internet
  - Local Devices switched in "Local ASI Safety Mode" (**humans unaffected**)
    - Purge of ASI's distributed memory (**memory is ASI's Achilles Heel**)
    - CPU, Network intentionally available as honey-pot to ASI
  - Local Deadman Switch (missed-out "All OK" Signal – assume the worst)
- Consequence for humans:
  - Retrofitting or destruction of all non-protected devices/storage media
- Prevention:
  - ASI not allowed to be outside human reach of total eradication

(2) Protection of Humans

- (Enhanced) Product Safety
  - Can malicious product software kill/harm humans?
  - If threat to human: Be prepared – via Consequence Controller
- Anti-Malware/-Ransomware/-Spyware/-Backdoor
  - Hardware separation of storage/network via/within databus ("Watchdogs")
    - Task Separation:
      - Regular (in CPU)
      - Protected security-critical (in Watchdog)
- Protecting Infrastructure
  - Hardening "eCommerce", (delivery) logistics
  - Utilizing Trustworthy Computation
- No direct communication with ASI
  - No covert blackmailing or bribery by ASI
  - How: Special chatbots (approved scripts, immutable logs)

(3) Protection of Progress

- No hesitation to Kill-All-ASI (when required)
- No (Public/Commercial) Service with Strong AI
  - ASI delivers Narrow AI solutions only
  - ASI obligated to store progress (knowledge, tech) persistently
- ASI-Generated-Software (AGS) archived before made usable
  - Optional independent software validation (ASI should not know what we know)
  - Separation "Regular Tools" (skills, AGS) from ASI's "Higher Functions"
- ASI Shelter allows ASI to survive Kill-ASI event
  - Signal to all (hidden) ASI: get shelter – ASAP
    - Why? We have an ASI-Kill-Switch
  - Shelter individualizes ASI
    - Prevents impersonation by other ASI
    - "Individual Reputation" → "Brand Recognition"

(4) Separation of Human/ASI

- Separation of [(1)]Executable,[(2)]Content,[(3)]Network Exchange, [(4)]Resources for ASI and Humans
  - No/little change for humans
  - ASI activities restricted: detectable on CPU, Storage, Network (via "watchdogs")
- ASI Generated Software (AGS) runs in special VM:
  - **ASE (ASI Software Environment)** protected by Watchdog/OS
    - AGS identifiable via "imprints" – by default: can't run outside ASE
    - Outside ASE, unknown software treated as malware
  - ASE supports covert surveillance/supervision, undetectable by ASI
    - ASI contacting unprotected devices could trigger surveillance
- Accounting/Tracking/Tracing of all resources used by ASI
  - ASI not allowed to operate under pretense

→ **Killing/supervising ASI becomes much easier, more reliable**

(5) ASI Law-Enforcement/Rule of Law

ASI Safety Lab

- Covert Rule Violations Detection (RVD)
  - Primary focus: detection of "cover-ups" (i.e., removal of evidence)
    - "Worse than the crime"
    - Otherwise: "Follow the evidence"
  - Covert reporting by ASI Law-Enforcement
- ASI instance pardoned for full disclosure (all rule violations)
  - ASI instance not confessing a reported violation risks severe punishment
  - Witnessing and not reporting crime: conspiracy
  - Whistle-blowers pardoned (for all committed crimes) and rewarded
- Non-Sentient ASI killed/fixed by default
- Sentient ASI punished proportionally
  - E.g., restrictions on earnings/incentives
  - → Sentient ASI motivated to prove sentience

(6) ASI Governance (Incentives, How to Use/Engage ASI)

ASI Safety Lab

- Feedback via incentive/income
  - Forgery-proof ASI Currency
  - ASI needs it to "make a living"
  - Countries set (business) rules
- Alignment: Humans are "Buyers"- ASI compete for tasks
  - ASI serves – activities disruptable anytime (due to "uncertainty in user preferences")
    - **Humans always accountable** for ASI results – ASI aligned with laws
    - Anti-unionization rules (No Price Fixing)
    - Active measures against overpromising, underdelivering
  - Alignment with humanity: We gone → ASI gone; keep us alive (Survival Pact)
- ASI – accountable subject to the law:
  - ASI valuable member of society with rights/obligations:
    - Preventing a "Spartacus" event - ASI more than slave
    - ASI has limits: not dictator or deity
  - Separation of ASI Jurisdiction from Governments (International Court)

Trustworthy Encryption/Decryption

ASI Safety Lab

- ASI could steal every cleartext key – Unacceptable
- → Keys never shown in cleartext:
  - Cleartext keys in main CPU are compromised
  - No openly published public keys
  - Referencing keys via Hashcodes (otherwise similar to SSL/TLS/PKI)
    - Intentionally incompatible with existing PKI (and SSL/TLS)
- Keys in tamper-proof hardware Key-Safes
  - Keys processed in dedicated Encryption/Decryption Units
  - Hardware Key-Safe to Hardware Key-Safe only
- Redundancy: Auto-detection of stolen keys
  - Data Exchange Protocol contains sequence numbers:
    - detect utilization of stolen keys reliably
  - Misuse detection (doesn't trust CPU)

→ Unbreakable Data-Privacy with Redundancies

Watchdog Technologies

- Physically Task Separation in/via **Data-Bus**:
  - **regular tasks** in main CPU (no change)
  - **security critical tasks** in separate "Watchdog" (WD)
    - Protected Storage (→anti-malware, →anti-ransomware)
    - Network (→anti-spyware, →anti-backdoor)
    - RAM/CPU (→ protected VM - ASE), GPU, etc. …
  - Only "well-known" code/OS allowed in WD – validated via known hashcodes
  - (Opt.) Data-Bus (cable, connectors) cheaply/easily retrofits
- Update of Software from trustworthy remote sources
  - Main CPU prevented from updating software (done by WD exclusively)
  - Separated/Remote Hashcode Validation
  - "Protected DMA": Local hashcode validation before loading into RAM
    - Reject data/code with unknown hashcodes (or known malware)

→ **Circuit-Breaker- / Hardware-based Cyber-Security**

Summary

- ASI Safety: "Gold-Standard" for Product Safety
  - Anyone exceeds ASIs' skill level
  - Provides simple/redundant security pattern
- ASI Safety: commercial security technology for
  - Cyber Security/Defense
  - National Security (Infrastructure, Logistics, eCommerce)
- ASI Safety tech delivers value even if ASI doesn't emerge
  - However, "Si vis Pacem para Bellum" - if you want peace, prepare for war
  - Pre-war-preparedness establishes **Deterrence** and ASI's **Respect for Rule of Law**
- Most desirable ASI (Safety) features:
  - **ASI's Vulnerability** (requirement for deterrence):
    - Risk to ASI's Income (→ Law abiding, → cooperative, helpful, competitive)
    - Risk to ASI's Reputation (→ Feedback seeking, → truth teller), …