# A Case for Trustless Computing[1]

*Rufo Guerreschi, Udit Dhawan and Roberto Gallo*

Trustless Computing Association

http://www.trustlesscomputing.org

Version 0.95 (10/10/2017)

## Abstract

*Recent revelations and reported security breaches have highlighted that current IT standard and certification processes are gravely inadequate in their ability to (a) enable citizens to reliably evaluate or have access to IT services that are coherent with their fundamental* ***civil rights****, (b) support adequately the implementation of public policies aimed at the defense of citizens' civil rights, critical assets, democratic sovereignty, and integrity and efficacy of targeted* ***cyber-investigations****, and (c) provide a suitable baseline for national and international regulation and certification of the most critical deterministic sub-systems of advanced and high-volume critical* ***AI systems*** *and projects, given their huge long-term societal implications.*

*Goals (a) and (b) have increasingly revealed themselves as interlinked, because the failure to provide (a) has been overwhelmingly due to efforts by powerful nations to retain cyber-investigation capabilities through endpoint cracking against digitally sophisticated criminals states and individuals. This has in turn prevented such endpoint cyber-investigation capabilities to achieve nearly the required levels of integrity of evidence to withstand court requirements, and the required resistance from external and internal abuse to foster the necessary international intelligence exchange in the face of grave terrorist threats.*

*We, therefore, propose the institution of a new standard setting and certification body, the Trustless Computing Foundation, and related initial constituent processes and high-level standards, or Paradigms. It will aim to reliably provide ordinary citizens access to affordable and user-friendly IT services, while offering levels of trustworthiness that are ultra-high and meaningfully-abiding to the EU Charter of Fundamental Rights, and reliably ensure "constitutional" lawful access; and provide the highest level of assurance certification for the most critical deterministic sub-systems of critical AI systems. Such standards will in time be compliant to EU, international and national privacy regulations, as well as, prospectively, compatible with leading high-assurance standards in IT communications, such as Common Criteria, SO-GIS, and leading and emerging critical infrastructure and critical AI standards.*

[1] This paper is a summary of an extensive detailed proposal being drafted by the Trustless Computing Initiative, and several speakers of the Free and Safe in Cyberspace event series. The full working draft can be found at www.trustless.ai/docs/tc_whitepaper.pdf

# Table of Contents

# 1. Introduction

*Constitutionally-meaningful* **confidentiality and integrity** of digital data, and the **preservation of effective cyber-investigation capabilities**, are not, as most believe, an "either/or", but a "both or neither”. In fact, neither digital confidentiality nor effective cyber-investigation capabilities are available today because nearly all IT[2] services - including cyber-investigation tools - can be compromised at scale through vulnerabilities that some criminal entities and powerful nations have directly implanted or indirectly sanctioned - by hugely financing the market for zero-day vulnerabilities, by deliberately subverting key lifecycle phases, by neither disclosing nor properly safeguarding found vulnerabilities, and by deliberately promoting broken standards.

While **perfect trustworthiness is impossible**, it is crucial to set a measurable level of target trustworthiness in confidentiality and integrity of IT services, which are sufficiently resistant to being compromised at scale, in order to substantially increase the levels of freedom and democracy in the society. Such levels should at least enable citizens to responsibly and effectively exercise their internet-connected communication civil rights, in accordance to international human rights agreements such as the UN Charter and the EU Charter of Fundamental Rights, except for remote voting in governmental elections and political primaries.

Ultimately, it is a “**both or neither**” challenge because, to a very large extent, the extreme technical, cyber-social and organizational *safeguards* that are needed to ensure *ultra*-high levels of assurance of communications, are the same safeguards that are needed to define innovative lawful access compliance protocols and certifications which will reduce the risks of widespread abuse of the civil rights of citizens AND of integrity of cyber investigations to levels that are (a) “**constitutionally**” acceptable, or (b) at the very least several times **less invasive** than current common and best practices.

We propose that the assurance of any IT service should not be assessed according to reputation or compliance of part of its critical components to **insufficiently comprehensive and self-referential** certification standards, as it is done today through the dominant “trusted computing” model. Rather it must be measured through a fine-grained continuous modeling and real-time transparent monitoring of all relevant and intrinsic technological and procedural constraints and all relevant organizational, economic, liability, legal and sociobehavioral disincentives, that might cause critically-involved individuals and organizations to perform unexpected compromising actions.

It is therefore necessary that “so called” **privacy-by-design** and **security-by-design** paradigms be brought to their ultimate conclusion, by requiring that IT services be `*trustless*`, i.e., **devoid of the need or assumption of unverified trust** in anyone or anything, except in quality of self-guaranteeing transparent and accountable organizational processes, that underlie all critical services and technology lifecycle and provisioning, and whose quality is recognizable by moderately informed and educated citizens.

To that end we propose the Trustless Computing Certification Body Foundation (or Trustless Computing Foundation) and its initial statutes, by-laws and high-level binding guidelines embedded in the Trustless Computing Paradigms (or “Paradigm”) to achieve and sustain actual and perceived levels of trustworthiness of IT systems that are today largely deemed impossible, inconceivable or uneconomical, and ensure its wide adoption and affordability by all.

[2] Information and Communications Technologies -- for practicality we use “IT” and “computing” interchangeably

# 2. Aims, Scope, Ambition

We define an *ultra-high assurance* computing service as one that can be confided in by the *users* to be resistant to persistent attempts by actors with high plausible deniability and very low actual accountability, with budgets in the range of tens of millions to compromise its lifecycle, and with budgets in the range of tens of thousands to persistently compromise single users, such as those associated with on-site, proximity-based user surveillance, or non-scalable remote endpoint techniques. For practical purposes, our scope will include only the security characteristics of confidentiality and integrity, of both data and metadata, and not of availability, initially.

We define, instead, a Trustless Computing IT service (or "TC service"), an *ultra-high assurance* computing service that, in addition to providing ultra-high levels of trustworthiness by the *user* towards a given IT service or experience, **also provides ultra-high levels of trustworthiness *by society* against abuse by the users of such services - and/or its publicly-available designs - to hide, plan or execute grave crimes**. Trustless Computing must therefore include and mandate compliance by service providers to "constitutional" - no more and no less - lawful-access to user data, without creating unacceptable risks to the privacy of users. Unacceptable risks will be defined, at a minimum, as risks that are higher than the best alternative that is commercially available.

Existing international standards and certifications do not currently provide for Trustless Computing levels of trustworthiness. Our aim is to create a new international standards setting and certification body for end-to-end IT services, and related lifecycles, that can reliably certify *ultra-high assurance computing* services and *Trustless Computing* services -- the **Foundation**. Together with the parallel establishment of Trustless Computing Consortium - comprised of technical, end-user and commercialization members - the Foundation will **create and sustain highly-competent and user-accountable** computing standards, a profitable ecosystem of independent service providers, and an active ethical hacking community, around the creation and decentralized evolution of the **world's most user-trustworthy general-purpose computing services platform, lifecycle, and standards**. The consortium initiative originates from the emerging Trustless Computing Consortium of the Open Media Cluster and its core partners[3]. The objectives of the Foundation and Consortium will be to:

(A) Realize standards for end-to-end computing services delivering **ultra-high levels of assurance** in privacy, authenticity and integrity that are compatible with a justified confidence in a meaningful protection from remote abuse of users' constitutional communication rights. These should allow users and service providers to achieve and sustain actual and perceived levels of trustworthiness of IT systems that are today largely deemed impossible, inconceivable or uneconomical, and to ensure its wide adoption by millions and its ultimate affordability to any citizen no later than three years, albeit through minimal initial features and performance -- these will initially mean to complement, and not replace, a typical user's everyday desktop or mobile computing.

(B) Enable any willing service provider to offer such computing services, creating a **highly-decentralized and participatory ecosystem** of organizations, whose technical and user-accountability effectiveness is highly-resilient to advanced persistent threats along short- and long-term changing technological, legislative and societal contexts. They will provide checks and balances among different user-accountable organizations, effective organization re-constituent processes, autonomous communities of self-provisioning users, competing service providers, and even competing standards organizations.

[3] https://www.openmediacluster.com/trustless-computing-consortium/

# 3. Cyber-Social Breakthroughs

Over the last few decades, the IT security community has made significant breakthroughs to enhance the security (in terms of confidentiality and integrity) and assurance of IT services and devices - from end-to-end encryption to blockchains[4], to anonymization networks, to formal verification in hardware and software designs[5]. However, most of our systems are still vulnerable to kinds of subversions that these technologies were *designed* to prevent, as is evident from recent revelations and reported security breaches, and significant increases in cybersecurity spendings and cost of cyber-thefts[6,7] over the last few years alone. The result is an ever-increasing, ever-looming threat to the privacy and confidentiality of digital communications and transactions for individuals, enterprises and governments alike, and to the integrity and efficacy of law enforcement and governance.

Several cybersecurity experts have continually raised their concerns towards this state of affairs. In 2013, Edward Snowden said[8], "*Encryption works. Properly implemented strong crypto systems are one of the few things that you can rely on. Unfortunately, endpoint security is so terrifically weak that NSA can frequently find ways around it"*. Bruce Schneier, one of the top cyber-security experts alive today, has been an active voice on the state of cybersecurity and has consistently emphasized on the exploitable vulnerabilities that arise from lack in oversight in the fabrication processes (supply chain attacks), to none to little verification and/or verifiability (primarily due to their complexity or lack of such models[9]) of software and hardware designs to governance[10].

In crux, today's (and tomorrow's) large scale IT services and devices **cannot** rely only on the technical aspect of cyber-security for trustworthiness, because all IT services, products and systems (user side as well as ones used by law enforcement agencies for cyber investigations), ultimately and necessarily, critically rely on at least some services during their development, and after initial sale or deployment (such as software upgrades) - that is, their entire lifecycle. This implies that the overall trustworthiness of an IT device or a service, in addition to its technical aspects, is inherently dependent on the cumulative trustworthiness of lifecycles of any and all of its critical technical and organizational components.

To fix this state of affairs, we have to begin by reconceptualizing cybersecurity and trustworthiness as exclusively a *cyber-social* problem, and not a technical one. Of course there are technical innovations that can affect cybersecurity but the likelihood of their occurrence and their availability to law-abiding actors rather than non law-abiding actors is all a matter of governance. We have to redefine cybersecurity as a **by-product** of the **intrinsic resilience, accountability and technical proficiency** of all components (such as algorithms, software, protocols, hardware), and organizational processes (such as supply chain, human processes, standard setting) that are critically involved in the entire life-cycle of producing and consuming an IT product. Once we do that, cybersecurity and trustworthiness are, in fact, a *governance* problem with a combination of technologies, regulations, economic (dis)incentives and social norms. Achieving trustworthiness by being ***trustless*** will require uncompromisingly applying best-of-breed "*zero trust*" social and technical paradigms from different fields, including:

[4] https://hackernoon.com/using-blockchain-technology-to-boost-cyber-security-19b6ef4e6898

[5] https://www.quantamagazine.org/formal-verification-creates-hacker-proof-code-20160920/

[6] https://www.csoonline.com/article/3210912/security/is-cybercrime-the-greatest-threat-to-every-company-in-the-world.html

[7]https://www.forbes.com/sites/stevemorgan/2016/01/17/cyber-crime-costs-projected-to-reach-2-trillion-by-2019/#30e0113c3a91

[8] http://www.businessinsider.com/edward-snowden-email-encryption-works-against-the-nsa-2013-6?IR=T

[9] https://www.schneier.com/blog/archives/2009/10/proving_a_compu.html

[10] http://listen.datasociety.net/security-privacy-hyper-connected-world/

(a) **cyber-social principles of highest-trustworthiness** military IT and civil aviation systems, such as secret sharing, threshold cryptography, blockchain and zero-knowledge protocols for cryptography and human process protocols,

(b) **citizen-witness, citizen jury and voting-booth** organizational procedures in democratic elections, and

(c) organizational constituent processes, and statutory architectures, aimed at **extreme transparency, user/citizen-accountability and technical-proficiency**

# 4. Trustless Computing

Governance is about constituent processes. The sustainability in time of the democratic and technical quality of such governance is ultimately wholly dependent on the foreseeable ability of the initial organizational statutes, and members of initial key governing boards, to maximize the chances of self-improvement, amidst the pressures of growth and success, because *"One cannot in the nature of things expect a little tree that has been turned into a club to put forth leaves*", said Martin Buber. In the Trustless Computing paradigm, the trustworthiness of any end-to-end IT service or experience will not be assessed according to organizational cognitive trust (reputation) and compliance to gravely incomplete and auto-referential certifications standards (e.g. Common Criteria, FIPS, Trusted Computing), as done today. Rather, cybersecurity will be assessed and certified as the level of trustworthiness that individuals and organizations critically-involved will not perform unexpected actions, and shall be derived from dynamically modeling all technological, procedural and statute cyber-social intrinsic constraints, and all organizational, economic, liability, legal and social disincentives, that are foreseeable at any given time. Trustless Computing paradigms are based on and derived from the following key concepts[11]:

(1) complete verifiability, extreme minimization and compartmentalization, and sufficiently-extreme verification relative to complexity of all critical hardware and software.

(2) extreme oversight, centered on offline citizen-witness and citizen-jury processes, of all critical technical and socio-technical components during their entire lifecycle, including critical hardware fabrication and server-room access, and allowing for "constitutional" lawful access requests.

(3) extremely technically-proficient and citizen-accountable IT assurance standards setting and certification governance.

## 4.1. Trustless Computing Paradigms

In this section we define the key paradigms that form the foundation of any TC service. These are intended to guide not only the establishment of a certification standard, but also to ensure and sustain a highly resilient and open ecosystems that are fully coherent with such standards. These are meant to be binding in nature in the sense that a compliant provider will need to respect them throughout the lifecycle of the service or the device to consistently maintain Trustless Computing certification. These are in no way meant to be exhaustive; they also provide a mechanism to make amends to the paradigms themselves.

A compliant Trustless Computing service by a provider will therefore be described as one which:

A. AIMS: *aims* at substantial **constitutionally-meaningful levels of actual and perceived trustworthiness** to the end-user of the confidentiality, anonymity, integrity and authenticity

[11] More details can be found in the full working draft on Trustless Computing -- www.trustless.ai/docs/tc_whitepaper.pdf

of data and metadata of his/her entire connected computing experience, and not mere substantial improvements;

B. SCOPE: *aims* to provide a user-friendly **supplement or "add-on"** to ordinary commercial mobile and desktop devices, rather than a replacement to them, with substantially or radically unprecedented levels of trustworthiness.

C. EXTENT: *comprehends* **all critical service components, meaning all hardware, software or organizational processes** involved during the entire lifecycle and supply-chain, at the endpoints, and in the overall architecture of midpoints relevant to the ensuring of metadata privacy; i.e. those whose possible vulnerabilities and critical weakness can NOT be protected against, at the highest-levels of trustworthiness, through compartmentation such as proven OpSec, OS, IC/SoC or CPU-level isolation techniques.

D. MEASURE: *assumes* that extremely **skilled attackers are willing to devote even tens of millions** of dollars to compromise the lifecycle or supply chain through legal and illegal subversion of all kinds, including economic pressures; and many tens of thousands to compromise of the individual end-user. Rate levels.

E. TRUSTLESS*ness*. *assumes* an **active and complete lack of trust in anyone or anything, except** in the intrinsic constraints and incentives against decisive attacks to all organizational processes critically involved in the entire lifecycle, from standard setting to fabrication oversight, as assessable by any moderately informed and educated citizen.

F. ORGANIZATIONS: *provides* extreme user **accountability, independence and technical proficiency of all organizational processes critically involved** in the computing service lifecycle and operation, including the certification body or bodies. Involves direct and exhaustive involvement of informed samples of citizens in the design and operational security oversight of all critical components.

G. CRYPTO: ***includes*** **only highly-redundant hardware and/or software cryptosystems whose protocols,** algorithms and implementations are either open, long-standing, standards-based and extensively verified and endorsed by recognized ethical security experts, albeit with lesser performance, and widely recognized for their post-quantum resistance levels, aiming at a migration to post-quantum cryptography in the next 5-10 years, including, but not limited to:

   a. **privacy-preserving** mechanisms such as zero-knowledge proofs for operations such authentication and integrity checks

   b. strong **decentralized** cryptographic protocols such as blockchains and threshold encryption

   c. emerging **homomorphic** encryption to allow for computation on encrypted data without converting user data to plaintext

H. AUDITABILITY 1. *integrates and develops* only **software and firmware whose source code and compiler allows for auditing without non-disclosure agreement** ("NDA"), and which is developed openly and publicly in all its iterations;

I. AUDITABILITY 2. *includes* **only critical hardware components whose firmware (and microcode) and full hardware designs are publicly auditable without NDA** at all times in open public structured format. In the case of processors, it will include code, hardware description source files (such as VHDL or Verilog files), Spin interpreter and similar, programming tools, and compilers;

J. AUDITABILITY 3: *allows* for complete **hardware fabrication and assembly auditability,** and extremely user-accountable and effective oversight, of all critical hardware components, in their critical manufacturing processes;

K. AUDITABILITY 4: *ensures* availability of **one or more mirror physical copy of the complete server-side hosting room setups** to enable easy independent testing by anyone, while being charged only the marginal cost of providing such access; in addition to all needed service devices at marginal production cost

L. *ACTUAL AUDIT. provides* **extreme levels of actual auditing relative to complexity**; i.e. levels of intensity, competency, and "expected altruism" of engineering and auditing efforts deployed, relative to complexity, for all critical software and hardware components, including through extreme software and hardware compartmentation;

M. LICENSE. *strongly **minimizes*** **the inclusion of non-Free Software**, including updatable and non-updatable firmware. Makes extensive reuse of existing Free/Open Source Software components – through extreme stripping down, hardening and re-writing. It strongly aims at realising the computing device with the least amount of non-free software and firmware in security-critical hardware components;

N. TRAINING. *includes* **effective and exhaustive first-time in-person training for users**, to ensure knowledge of basic operational security (OpSec) and the risk management for self and others. This, in addition to the absence of externally-exposed ports and presence of effective tampering detection on the end-user devices, aims to provide most or all the benefits of *remote attestation*, which is not permitted due to its significant risks. Users must be able to fully reprogram the device using an internal port after triggering the tampering detection mechanism;

O. IP TERMS: *includes* **only technologies and innovations with clear and low long-term royalties** - from patenting and licensing fees - to prevent undue intellectual property right holders' pressures, lock-ins, patent vetoes, and ensure an open platform with sustainably low costs, affordable to most western citizens.

P. LEGAL*: ensures* that current cyber-security **legislations and state agencies practices** in the country of origin and/or localization of user, provider, assembly facilities, foundry - and other critical process involved - are consistent with a constitutional/lawful and feasible compliance with standards; in regards to surveillance, mandatory encryption key disclosure, crypto exports, liability, and other relevant legislations.

Q. ASSEMBLY. *provides* one or more dedicated crowded urban street-level glass-walled spaces where devices are publicly assembled, verified, flashed, and transferred to their users. It will be subject to 24/7 high-trustworthiness live streaming oversight, and monitoring.

R. LIABILITY: *includes* an extreme level of **cumulative liability, contractual/economic and legal,** for all individuals and organizations critically involved for not strictly following procedures or willingly compromising the life-cycle.

S. OPEN ECOSYSTEM. *involves* participants to an initial open R&D Consortium, which will set out to build the first certified service, that **commit to terms that ensures very-high resilience to the openness of the ecosystem** and its resistance to economic pressures, including: (a) through such consortium, offer only certified services; (b) state clear, perpetual and very-low (or null) royalties to all the IP they integrated and developed in the services they offer jointly or independently.

T. SERVER-SIDE. *may* provide privacy-sensitive server-side services on condition that they are provided with very extreme safeguards from abuse, at the following conditions:

  a. only through extremely technically-effective, citizen-accountable and transparent safeguards, whose effectiveness is reliant on highly-resilient **citizen-witness**-supported on-site physical access management organizational processes of involved hosting facilities, similar to those that govern high-standard paper-based ballot box voting. These include the ability and strong obligation of those randomly-selected citizen witnesses to prevent attempts to procedural violation by anyone, by reliably and promptly causing either such services' termination and secure erasing of sensitive data, or their immediate or deferred transfer to an alternative safe hosting room. Key operations of the system must not depend on the availability of the hosting room;

  b. only if both the provider and the hosting facility are located in nations where legislation or known practices, do NOT make it illegal - and with less than negligible consequences - to withhold access to warrant-based or state-security-based government requests. Terms of service and operational procedures must in fact clearly exclude compliance to any government request for personal data of users. When and if laws are changed that make it illegal, then the Provider must give a choice to each individual user to either (a) agree to transfer such services to other nation where it is legal, or (b) turn off such server-side services. Providers that are governmental agencies, civilian or military, and offer service to public employees are exempt, transparently to their users, from the requirement of this clause.

U. HOSTING ROOMS. *deploys* only **TC-compliant devices** as for any *critical* function, where **remote admin access is disabled**; involves state-of-the-art public **video streaming and recording**, and is located at street level in busy urban street with large glass fronts, to increase perceived (and actual) social control;

  a. on-site access by anyone is conditional on the physical presence and approval of a minimum number of citizen witnesses;

  b. enables citizen-witnesses to launch a "scorched earth procedure"[12], with plausible deniability, which physically burns all data;

  c. may rely on an additional layer of safeguard by allowing a set of users located in a different Member State and/or randomly selected users to act as "remote witnesses", as an additional layer of oversight, using secret-sharing and threshold approval/cryptographic techniques;

  d. will maintain one (or more) complete replicas of the complete infrastructure which will be publicly available for complete audit tests;

  e. sets intrinsic technological limits to the maximum number of users and percentage of total users whose personal data or keys may be recovered within a given time frame;

  f. may make use of additional safeguards, such as protection via *implicitly learned passcodes*, that cannot be revealed explicitly by the user and may increase the plausible deniability in case of emergencies, and the related "scorched earth procedure";

[12] http://www.usatoday.com/story/money/columnist/rieder/2013/08/12/reider-nsa-snooping-collateral-damage/2642557/

V. FABRICATION. *ensures* that the requested hardware is all produced in one continuous batch in a short time span (a few days or weeks), as is typical anyway, and

   a. adds a minimum number of "user witnesses", in a role of active oversight witnesses 24/7.

   b. chooses to produce critical ICs (such as CPU, SoC, memory, etc) at foundries with older technologies, simpler processes, and less third-party IP obstacles than today's Asian mega fabs, that allow the technicians and witnesses to publicly and completely document the process with videos, photos and more.

   c. uses equipment and sensors, to be applied to the chosen foundries, that as much as possible not require direct interventions or disruption of the foundry equipment and facilities, but just rely on setting up an additional overlay of sensing equipment, and on getting copy of the existing quality control sensor feeds.

# 5. Trustless Computing and Backdoors

## 5.1. Why the proposed state-mandated backdoors would be useless for public safety and extremely dangerous for citizens' security and privacy

In recent times, several state authorities and intelligence agencies have proposed to solve the "going dark" problem by mandating some kind of backdoors into all IT systems. The FBI has more specifically proposed a "*legislation that will assure that when we get the appropriate court order . . . companies . . . served . . . have the capability and the capacity to respond"*[13], while the NSA has been generically referring to organizational or technical safeguards ensuring backdoor access authorization approval by multiple state agencies.

Since the 1990s, in the legitimate pursuit of extending the lawful access state had to all other means of communications, to IP Systems, many national legislative proposals for *exceptional access* (or state backdoors) have aimed to mandate technical systems that enable covert remote access into all IT server-side services or user-side devices - sold or introduced in the country - by lawful agencies. These proposals, if enacted in laws or treaties, would have a decisive negative impact on both citizens' privacy and for public safety, for the following main reasons:

- The most advanced public security agencies have had, and likely will continue to have, the continuous capability to break into nearly all endpoints, at nearly all times, for targeted surveillance; they have in fact needed to resort to such comprehensive capability precisely from the 1990s when unbreakable encryption became popularly available.
- Given the enormous complexity and diversity of IT systems and providers, it would be both highly expensive and practically impossible to verify and certify implementations that are sufficiently trustworthy.
- Legislative and public security branches of government have proven deeply and repeatedly their lack of competency in architecting technical standards and oversight processes to reasonably limit their abuse.
- Criminals could still surreptitiously fabricate, modify or import - or use while abroad - IT systems without such built-in access, and could still pre-encrypt messages externally to the device or use other means, such as steganography, to communicate covertly over such IT systems.

[13] https://www.cnet.com/news/fbi-looking-at-law-making-web-sites-wiretap-ready-director-says/

### 5.2. How the Trustless Computing lawful access compliance mechanisms differ from state-mandated backdoors

Our proposal for extreme safeguards certification for voluntary compliance to lawful access requests, by ultra-high assurance IT providers, is out of the declared scope of the foundational report, "*The Risks of Key Recovery, Key Escrow, and Trusted Third-Party Encryption*" by some of the world's top IT security experts, because it is:

1. Not mandated by the state. Instead, it is a voluntary practice, i.e. in addition of current legal requirements - by certified ultra-high assurance IT providers, certified by an international certification body, and only in selected jurisdictions where laws and practice allow for the provider, or others they delegate, to safely exercise discretion on the basis of constitutionality of the lawful access request;
2. Not regulated, designed, standardized **or** certified by the state. Such functions would be managed by a *trusted third-party,* in the form of an extremely technically-proficient and citizen-accountable international standard setting and certification body, the Trustless Computing Foundation, and by temporary organizational entities made of groups of randomly sampled citizen-jurors and citizen-witnesses, tightly regulated by such body;
3. Not universal for all IT systems. It is reserved only to ultra-high assurance IT devices and services, such as Trustless Computing for wide market use, which can truly be expected to be beyond the targeted or large-scale exploitation capability of a large number of democratic nation states' cyber-investigation agencies.

Nonetheless, the above "*The Risks of Key Recovery, Key Escrow, and Trusted Third-Party Encryption*." report and its subsequent "*Keys under doormats*" often move beyond their main stated objective, to preemptively criticize extensively any conceivable way by which compliance to state lawful access could be voluntarily offered by providers without creating *additional* unacceptable risks of abuse to the users.

# 6. Radical Mitigation of Abuse of Publicly-Available Ultra-high Assurance IT Designs

The public verifiability of the source designs of every critical software & hardware component as prescribed by the Trustless Computing paradigms could appear to potentially enable malevolent actors to fabricate their own devices beyond the capability of interception by even the most power intelligence. In fact, several large non-EU non-NATO non-allied countries already have all the capabilities to build systems to the Trustless trustworthiness levels, and could make it available to malevolent actors. Nonetheless, we have carefully concocted preliminary definitions of safeguards to sufficiently and radically mitigate such a threat.

In theory, smaller potentially malevolent states or groups, by contrast, in order to achieve and sustain the Trustless levels of assurance, would need to have an extreme control of a suitable semiconductor foundry, because, as US Defense Science Board said back in 2005, *"Trust cannot be added to integrated circuits after fabrication*". The dramatic increase in the complexity of critical HW fabrication and design processes[14] makes avoiding the insertion of an undetectable critical vulnerability throughout the supply chain and lifecycle an easy task. Furthermore, even a small foundry, by current global standards, is a highly complex operation with a staff of over 1000 and

[14] See this in depth analysis by Prof Villasenor: http://www.brookings.edu/research/papers/2013/11/4-securing-electronics-supply-chain-against-intentionally-compromised-hardware-villasenor

typically 800 or more discrete fabrication processes over several weeks, including dozens of critical ones where an error or malicious modification, can not be detected afterwards. Provisions will be set in the HW/SW architecture to ensure that Trustless Computing compliant endpoint devices cannot be produced in smaller prototyping labs, mainly through the use of IP cores tied to specific, capital intensive fabrication processes, naturally not available on mini-scale prototyping fabrication facilities and foundries.

Furthermore, fabrication oversight procedures are needed because of the grave and real risk that hardware vulnerabilities may be introduced by some entity during the manufacturing process, and inadequacy of current fabrication standards. Such introduction, if performed in critical fabrications phases, cannot be ascertained afterwards. At first, it would appear that building a chip manufacturing plant would be the best way to provide the highest security of the chip manufacturing process. However, at a cost of hundreds of millions of dollars, for very old technology, to billions of dollars, for the latest, such costs are not only prohibitive but of very little use since, even though such plant may be located in the same nation where the Trustless service is offered, the problem of verifying and overseeing the process remains almost completely intact. Therefore, even if there was a budget of over $100M available to ensure hardware security, the best way to spend such budget would be in oversight procedures and technologies rather than manufacturing, provided that the necessary foundry access is granted.

In the rare case in which the malevolent entity might attempt to enter into agreements with suitable foundries to build such systems, state intelligence can easily make sure to either prevent it or, better yet, insert vulnerabilities in their fabrication or design processes to acquire in the future extremely valuable intelligence.

To the extent that the above mentioned safeguards may prove to be insufficient to adequately prevent such risks, the Trustless Computing Foundation may explore the possibility that a subset of the hardware designs - as opposed to all other critical technical components - may not be made public, but subject to multiple redundant verifications which involve direct oversight processes involving both randomly sampled citizens and elected officials, under suitably controlled environments.

# 7. Trustless Computing, Artificial Intelligence and the Future of Humanity

It is becoming increasingly clear that the balance of power in society, and the prospects of well being for the human race, are and will be increasingly dominated by the dynamics of formal and surreptitious control over the most advanced artificial intelligence systems and projects. Rapid developments in AI specific components and applications, theoretical research advances, high-profile acquisitions from hegemonic global IT giants, and heartfelt declarations about the dangers of future AI advances from leading global scientists and entrepreneurs, have brought AI to the fore as both (a) the key to private and public economic dominance in IT, and other sectors, in the short-to-medium term, as well as (b) the leading long-term existential risk (and opportunity) for humanity, due to the likely-inevitable “machine-intelligence explosion”, or singularity, once an AI project will reach or approach human-level general intelligence, at least in its capacity to improve itself.

A recent survey of AI experts[15] estimates that there is a 50% chance of achieving human-level general intelligence by 2040-2050, while not excluding significant possibilities that it could be reached sooner. Such estimates may even be biased towards later dates because, (a) there is an intrinsic interest in those that are by far the largest investors in AI – global IT giants and USG – to

[15] https://nickbostrom.com/papers/survey.pdf

avoid risking a major public opinion backlash on AI that could curtail their grand solo plans; (b) it is plausible or even probable that substantial advancements in AI capabilities and programs may have already happened but have successfully kept hidden for many years and decades, even while involving large numbers of people; as it has happened for surveillance programs and technologies of NSA and Five Eyes countries[16].

Stephen Hawking summarised it most clearly when he said[17], "*Whereas the short-term impact of AI depends on who controls it, the long-term impact depends on whether it can be controlled at all*". Control relies on IT assurance to ensure that those who control AI formally coincides with those who does so in practice, through hacking. It is unclear at this stage if *formal* control, in both the short- or the long-term, will have more influence on the nature of such AI systems than the *informal* control, i.e. the control exercised by those that have and will have sustained and undetected access to the most critical vulnerabilities of such systems.

In order to substantially reduce these enormous pressures, it is crucial to find ways by which sufficiently-extreme level of AI systems user-trustworthiness can be achieved, while at the same time transparently enabling due legal process cyber-investigation and crime prevention. Cyber-investigation capability may be crucial to investigate some criminal activities aimed at jeopardizing AI safety efforts**.** The solution to such dichotomy, proposed by Trustless Computing, would reduce the level of pressure by states to subvert secure high-assurance IT systems in general, and possibly – through mandatory or voluntary standards international lawful access standards – improve the ability of humanity to conduct cyber-investigations on the most advanced private and public AI R&D programs.

Trustless Computing paradigm, can become a crucial and fundamental element to increase the trustworthiness of the advanced narrow artificial intelligence systems (robots, self-driving cars, drones), and upcoming general artificial intelligence systems, by increasing the trustworthiness of their most critical *deterministic* endpoints and sub-systems by orders of magnitude. The dire short-term societal need and market demand for radically more trustworthy IT systems for citizens' privacy and security, and societal critical assets protection, can align – in a grand international vision – with the medium-term market demand and opportunity for large-scale ecosystems capable of producing AI systems that will be high-performing, low-cost and still provide adequately-extreme levels of security for AI critical scenarios.

Some may argue why extreme IT security to support AI safety is needed now if its consequences may be far away. One clear and imminent danger is posed by self-driving and autonomous vehicles (aerial and terrestrial) – which utilize increasingly wider narrow AI systems – and the ease with which they can be "weaponized" at scale. Hijacking the control of a large number of drones or vehicles could potentially cause hundreds of death or more, or cause hardly attributable hacks that can cause grave unjustified military confrontations.

Ideally, in our view, an international IT and AI assurance standard setting and certification body governance - with extreme technical proficiency and citizen accountability, as per our proposed body - would exercise effective formal and informal control on all known large private and public advanced projects to ensure both safety and humanity values alignment or, better even, guide

[16] https://www.privacyinternational.org/node/51

[17] http://www.independent.co.uk/news/science/stephen-hawking-transcendence-looks-at-the-implications-of-artificial-intelligence-but-are-we-taking-9313474.html

extremely well-founded international democratic nations' projects to develop "friendly AI", before "unfriendly AI" projects reach human-level general intelligence.

# 8. Manifesto of Trustless Computing v 1.0

Meaningful personal cybersecurity and effective cyber-investigation capabilities are not an "either or" but a "both or neither".

Neither are available today because nearly all communications IT systems, including cyber-investigation tools, are scalably compromisable through vulnerabilities that a few powerful nations have directly implanted or sanctioned, by hugely financing the zero day market, by deliberate strategic subversion of key IT lifecycles, by non disclosing found vulnerabilities and by deliberately promoting broken standards. This state of affairs is inevitable for nearly all current systems, even high assurance ones, because their technical and lifecycle complexity is by orders of magnitude beyond any sufficient verifiability, no matter what budget.

To a very large extent the extreme technical, cyber-social and organizational safeguards that are needed to ensure ultra-high levels of assurance of communications, are the same safeguards needed by lawful access compliance schemes that will reduce to acceptable levels the risks of abuse of the civil rights of citizens or the mission of national security agencies.

It is not inevitable, on the other hand, for IT systems, services and lifecycle that would certifiably implement **extreme levels of transparency, accountability, oversight, and security review relative to the complexity** of all critically-involved technologies and processes; from CPU design to fabrication oversight, from hosting facilities access management to standard setting governance.

Extreme compartmentalization and minimization in features and system complexity of hardware and software will allow unprecedented and consistently-extreme levels of security review relative to the complexity of ALL software, firmware, hardware and processes - including hardware design and fabrication, and hosting room management processes - critically involved in a Trustless Computing service, and its lifecycle. Open low-level compliant computing bases will ensure wide uptake..

Meaningful digital confidentiality and integrity, ultimately, are not a product, nor a service or a process, but the by-product of the relevant organizational and human process that are *critically*-involved in fruition, provisioning and lifecycle of a given IT service or experience. It is therefore critical that "so called" *privacy-by-design* and *security-by-design* paradigms be brought to their ultimate conclusion, by requiring that IT services be trust-free, i.e. devoid of the need of trust in anyone or anything, except in quality of self-guaranteeing transparent and accountable organizational processes, that underlie all critical service and technology lifecycle and provisioning, whose quality is recognizable by moderately informed and educated citizens.

According to the Trustless Computing paradigms, the assurance of any IT service will not be assessed according to reputation or compliance of part of its critical components to insufficiently comprehensive and self-referential certification standards, as it is done today through the dominant "trusted computing model". Rather it will be measured through a fine-grained continuous modeling and real-time transparent monitoring of all relevant technological and procedural intrinsic constraints and all relevant organizational, economic, liability, legal and social behavioral disincentives, that might cause individuals and organizations critically-involved to perform unexpected compromising actions.

# 9. Conclusion

Redefining trustworthiness in our current IT services as a cyber-social problem will allow us to account for the previously unrecognized and ignored aspects of cybersecurity and enable us to measure the trustworthiness in terms of the intrinsic resilience, accountability and technical proficiency of all components (such as algorithms, software, protocols, hardware), and organizational processes (such as supply chain, human processes, standard setting) that are critically involved in the entire life-cycle of producing and consuming an IT product. In this paper we proposed the formation of a standard setting and certification body with its initial statutes and guidelines to create an ecosystem of ultra-high assurance IT service providers and services that will significantly increase the trustworthiness of these services towards the users and towards the society.

# 10. Acknowledgments

This document was authored by **Rufo Guerreschi** (Executive Director of Trustless Computing Association, earlier called Open Media Cluster), **Udit Dhawan** (Founder and CTO at TRUSTLESS.AI) and **Roberto Gallo** (Founder and Security Tech Lead, TRUSTLESS.AI).

This Whitepaper Summary is a summary of the working draft of a full and formal Whitepaper of over 50 pages on a proposal for Trustless Computing Certification Body which can be accessed at www.trustless.ai/docs/tc_whitepaper.pdf, which also include detailed authors and contributors profiles. The content of this summary and of the mentioned full whitepaper has greatly benefited from discussions and contributions over 2 years of discussion in the Free and Safe in Cyberspace event series and the Trustless Computing Consortium R&D proposals, in particular with Bart Preneel (Ku Leuven), Jovan Golic (EIT Digital), Stefan Schuster (Tecnalia), Yvo Desmedt (U.Texas), Richard Stallman, Peter De Kostic (LIBE Committee Secretariat), Melle Van Den Berg (CapGemini), Michael Sieber (EDA Head of INformation Superiority). We took inspiration from writings by Steven Bellovin and Bruce Schneier, as well as all the advisors and partners of the *Trustless Computing* initiatives.

# 11. Further Reading

[1] Trustless -- TRUSTLESS socio-technical systems for ultra-high assurance ICT certifications, and a compliant open target architecture, life-cycle and ecosystem, for critical societal use case and consumer adoption (link)

[2] Facilitating Holistic ICT Certification (link)

[3] Trustless H2020 DS-01 RIA 2016 Memorandum of Understanding (link)