Engineering permanence in finite systems

The man-machine integration era (MMIE) is marked by sensor ubiquity, whose readings map human beings to finite numbers. These numbers processed by continuously changing, optimizing/learning, finite precision, closed loop, distributed systems are used to drive decisions such as insurance rates, prison sentencing, health care allocations and probation guidelines. Optimization and system parameter tuning is increasingly left to machine learning and applied AI. One challenge we face is thus: Ensuring the indelibility, the *permanence*, the *infinite value of human beings* as optimization-resistant invariants in such system environments. In this challenge paper, we propose developing safeguards, specifically working towards a 'deontological imprimatur' architecture embedding resilient representations of human beings.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2454v2 | CC BY 4.0 Open Access | rec: 12 Nov 2016, publ:

Engineering Permanence in Finite Systems

Daniel Bilar
Information Systems, Norwich University
158 Harmon Drive, Northfield, Vermont 05663, USA
Email: dbilar@norwich.edu

Abstract

The man-machine integration era (MMIE) is marked by sensor ubiquity, whose readings map human beings to finite numbers. These numbers—processed by continuously changing, optimizing/learning, finite precision, closed loop, distributed systems—are used to drive decisions such as insurance rates, prison sentencing, health care allocations and probation guidelines. Optimization and system parameter tuning is increasingly left to machine learning and applied AI. One challenge we face is thus: Ensuring the indelibility, the *permanence*, the *infinite value of human beings* as optimization-resistant invariants in such system environments. In this challenge paper, we propose developing safeguards, specifically working towards a 'deontological imprimatur' architecture embedding resilient representations of human beings.

Motivation

We motivate our exposition with the story of Thompson's fascinating 1996 experiment (Thompson 1997). His goal was to use genetic algorithms (GA, a set of optimization methods) to evolve a 10*10 cell circuit on a 64*64 cell FPGA (a configurable chip with cells consisting of transistors) that could distinguish between a 1 kHz and a 10 kHz sound wave. The circuit was *unclocked*, hence the GA was not evolving a digital system, but an analog continuous-time dynamical system of transistors (with input period five orders of magnitude longer than input to output signal propagation delay). The solution the GA found after 2-3 weeks had surprising properties: Certain FPGA cells outside the 10*10 solution circuit—with no connected wire path to influence the circuit—could not be removed without negatively affecting the solution. This meant that the GA included unexpected properties of the FPGA physical substrate, EM coupling or the power supply in its search space. Additionally, the solution was *non-transferable*, neither to other patches, nor other nominally identical FGPAs. It is thus not too far a stretch to imagine AI 'reward hacking'(Amodei et al. 2016) MMIE systems leading to different outcomes in testing or simulations versus operational settings.

Commensurability: Pitfall of Canonicalization

Optimization of MMIE systems will likely drive towards canonicalization of 'value'. When a human being is mapped to vectors of finite numbers, a incommensurable measure is effectively made commensurable. Commensurability allows for weighted utilitarian calculi; one example is Bentham's 'Greatest Good For Greatest Number'. When such calculi are used in optimization frameworks such as resource allocations, inhumane solutions—those that sacrifice the well-being or life of human beings for the 'greater' benefit of machine artifacts, or performance indices such as a greater energy efficiency—must be avoided, or at least readily identified. This is not as straightforward as it appears. Asimov's 1958 story "All the Troubles of the World" is a perfect example how readily a data-driven optimizing entity can seemingly innocuously work towards a hidden, catastrophic goal (Asimov 1959).

Infinite Value of a Human Being

Though the numeric range of digital finite precision systems seems daunting, notions of infinity are handled poorly. To wit, in IEEE754 both a very large and very small value are defined as $\pm\infty$. Thus, a very large number (say 1.1897×10^{4932} in quadruple precision 128 bit IEEE754 format) is still not qualitatively different from 27 in the Cantor Hierarchy of Infinity sense. One may attempt to mark human records in a system as undeletable within the system and a fixed rule "Never delete these records, no matter what benefits may accrue". But what happens when that system becomes part of a larger system, or is superseded by a copy without this restriction? Or the system learns how to transduce Rowhammer-style (Gruss, Maurice, and Mangard 2015)? How can we avoid data-commodified human elements, or conversely, the reification of a machine algorithm within the overall framework of a continuously optimizing environment?

Immortal Code, Data, Computations

Immortality in systems must in some fundamental manner resist legacy code refactoring approaches. This may be achieved by violating assumptions, by coercion, or by incentives (Feathers 2004)(Bilar 2010)(Anthes 2010).
One assumption-violating mechanism is *constant migration*.

To evade scanners, HBGary (a now-defunct security firm) proposed an assembly rootkit ("12 Monkeys/Magenta") that would inject itself into and rove rōnin-stylethrough processes while not associated with any identifiable object—no file, named data structure, device driver, process, thread, or module (Longpre 2011)(Anderson 2011).

One coercion mechanism is *abstraction opaqueness*. Windows *Win32k.sys* GUI subsystem is the oldest Windows OS component. In spite of a demonstrated porous attack surface, it still ships in Windows 10+, as a pre-Windows 3.1 legacy. Among other things it supports is Lotus 1-2-3 from 1983 (Mandt 2011).

Representation lock-in is a hybrid coercion/incentivization mechanism. IBM's mainframe 64-bit z/Arch (z/OS) architecture is largely backwards compatible with its ancient 1960s System/360 predecessor (Stephens 2014). IBM does not use the common ASCII standard, but rather an Extended Binary Coded Decimal Interchange Code (EBCDIC) character encoding. When mainframes served as standalone bastion systems (up and until the 1990s), this was fine. Nowadays, however, z/OS must communicate to and with everything—among other translations, the ASCII-to/from-EBCDIC translations are a major headache. Everything from REST interface parsing, to SSH connections, to simple file transfers have to grapple with this. Another complication is continuing support for the 1960s CKD (Count Key Data) disk architecture. z/OS can only use CKD formats. In the 1970s, IBM did try to move forward with fixed block architecture (FBA) for mainframes, but customers balked. They were accustomed to CKD, so IBM discontinued FBA format disks. To this day, IBM's direct-access storage device (dasd) must perform CKD emulation on top of FBA-type disks and arrays. In fact, CKD disks are not sold anymore, just FBA devices with controllers that perform CKD emulation (IBM)(McDaid 2012).

Embedding 'Immortality' into Finite Systems

We propose working towards a general safeguard architecture against human-endangering actions in MMIE systems. We maintain that representation of humans as *resilient, persistent information* is key to such an architecture. To this end, we posit that a mechanism inducing such a representation as a (1) *deontological imprimatur* is required. Such an imprimatur cannot be static, but (2) generative; must be (3) compulsorily enforceable; and have its secrets (4) provable, but (5) hidden.

By (1) 'deontological imprimatur' we mean a sine qua non condition that the system must cease to (usefully) function if the information were removed/disabled. In the parlance of a different domain and time, the conceptual equivalent of *élan vital* or a soul vivifying a body. Generativity (2) requires this information be refreshed at specific short intervals to be useful. Were one able to freeze this information statically in time, it could then be captured, removed from the system, and replayed as a hollow simulation. This would not meet our goal of a resilient, permanent embedding. Hence this 'soul' information needs to be dynamically generated anew—changing in some respects, yet unchanging in others.

For (3)-(5) we envision leveraging several mechanisms: Human representation (3) encoded as keys over a large number of entangled states (Yoshikawa et al. 2016) with (4) nonlocal multiplayer games (eg CHSH (Winter 2010)) and (5) zero-knowledge proof protocols within a HoTT formulation of closed-loop cybernetic control systems (Baez and Erbele 2015)(Aaronson 2016). We look forward to fleshing out further details with interested parties at the AICS workshop.

References

Aaronson, S. 2016. The Complexity of Quantum States and Transformations: From Quantum Money to Black Holes. In *Lecture notes for the 28th McGill Invitational Workshop on Computational Complexity*.

Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P.; Schulman, J.; and Mané, D. 2016. Concrete Problems in AI Safety.

Anderson, N. 2011. Black ops : how HBGary wrote backdoors for the government.

Anthes, G. 2010. Mechanism design meets computer science. *Communications of the ACM* 53(8):11.

Asimov, I. 1959. ALL THE TROUBLES OF THE WORLD. In *Nine Tomorrows*.

Baez, J., and Erbele, J. 2015. Categories in control. *Theory and Applications of Categories*.

Bilar, D. 2010. Degradation and Subversion through Subsystem Attacks. *IEEE Security & Privacy Magazine* 8(4):70–73.

Feathers, M. 2004. Working Effectively with Legacy Code.

Gruss, D.; Maurice, C.; and Mangard, S. 2015. Rowhammer.js: A Remote Software-Induced Fault Attack in JavaScript. *arXiv:1507.06955v1* 2016.

IBM. IBM ASCII to EBCDIC conversion.

Longpre, P. 2011. HBGary Rootkits: Catch Me If You Can!

Mandt, T. 2011. Kernel Attacks through User- Mode Callbacks. In *BlackHat*.

McDaid, F. 2012. (E)CKD VS SAN - The Scottish Mainframe Users' Group. In *SMUG*.

Stephens, D. 2014. Are IBM Mainframes Really Backward Compatible? - LongEx Mainframe Quarterly.

Thompson, A. 1997. An evolved circuit, intrinsic in silicon, entwined with physics. Springer Berlin Heidelberg. 390–405.

Winter, A. 2010. Quantum mechanics: The usefulness of uselessness. *Nature* 466(7310):1053–1054.

Yoshikawa, J.-i.; Yokoyama, S.; Kaji, T.; Sornphiphatphong, C.; Shiozawa, Y.; Makino, K.; and Furusawa, A. 2016. Invited Article: Generation of one-million-mode continuous-variable cluster state by unlimited time-domain multiplexing. *APL Photonics* 1(6):060801.