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 challenges we face is thus: Ensuring the indelibility, the *permanence*, the *infinite value of human beings* as optimization-resistant invariants in such system environments.**

I. 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, an incommensurable measure is 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 allocation, 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 smoother economy, CO_2 reduction, greater energy efficiency, roadblocks to refactoring etc —must be avoided, or at least readily identified.

This is not as straightforward as it appears: One may 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 But what happens when that system becomes part of a larger system, or is superseded by a copy without this restriction?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 (from JavaScript) [1]? How do we avoid (in Marxist terms) data-commodified human elements, or conversely, the reification of a machine algorithm within the overall framework of a continuously optimizing environment?

II. Immortal Code, Data, Computations

Immortality in systems must in some fundamental manner resist legacy code refactoring approaches. This may be achieved by violating assumptions, coercion, and incentives [2]–[4].

One assumption-violating mechanism is *constant migration*. To evade scanners, HBGary's proposed "12 Monkeys (Magenta)" assembly rootkit injected itself into and roved rōnin (浪人)-style through processes while not associated with any identifiable object; no file, named data structure, device driver, process, thread, or module [5], [6].

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. It supports among other things Lotus 1-2-3 from 1983. To a lesser extent, this is also true of IBM's JES2 job scheduler [7]–[9].

Yet another is *representation lock-in*. Communicating with IBM mainframe's z/OS EBCDIC encoding constitutes an ASCII conversion, and IBM CKD disk format (via ubiquitous FBA) an Inception incentivization nightmare, respectively [10]–[12].

In our view, a mixed mechanism inducing a *deontological imprimatur* is required. Such an imprimatur cannot be static, but generative; must be compulsory enforceable, and have its secrets provable, but hidden. We envision human beings encoded as keys in HoTT formulations of closed-loop cybernetic control systems over keyed entangled states, with non-local multiplayer games and zero-knowledge protocols [13]–[15]. This scheme was inspired by perceiving Quantum Music [16].

III. Past is Prologue

Asimov's 1958 story "All the Troubles of the World" delineates how readily a data-driven optimizing entity can seemingly innocuously work towards a hidden, catastrophic goal [17]. Thompson's fascinating 1996 experiment also serves as a cautionary tale [18]. 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. 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-transferrable*, neither to other patches, nor other nominally identical FGPAs. We can thus realistically imagine AI 'reward hacking' [19] MMIE systems (in conjunction with opaque signals) leading to different outcomes in testing or simulations versus operational settings. It is paramount we find ways to ensure that human beings are not insidiously optimized away.

This abstract is part of a series discussing MMIE issues [20].

REFERENCES

[1] D. Gruss, C. Maurice, and S. Mangard, "Rowhammer.js: A Remote Software-Induced Fault Attack in JavaScript," *arXiv:1507.06955v1*, vol. 2016, 2015. [Online]. Available: http://arxiv.org/abs/1507.06955

[2] M. Feathers, "Working Effectively with Legacy Code," 2004. [Online]. Available: http://dl.acm.org/citation.cfm?id=1050933

[3] D. Bilar, "Degradation and Subversion through Subsystem Attacks," *IEEE Security & Privacy Magazine*, vol. 8, no. 4, pp. 70–73, 7 2010. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5523869

[4] G. Anthes, "Mechanism design meets computer science," *Communications of the ACM*, vol. 53, no. 8, p. 11, 8 2010. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1787234.1787240

[5] P. Longpre, "HBGary Rootkits: Catch Me If You Can!" 2011. [Online]. Available: http://www.infosecisland.com/blogview/12678-HBGary-Rootkits-Catch-Me-If-You-Can.html

[6] N. Anderson, "Black ops : how HBGary wrote backdoors for the government," 2011. [Online]. Available: http://arstechnica.com/tech-policy/2011/02/black-ops-how-hbgary-wrote-backdoors-and-rootkits-for-the-government/5/

[7] A. Ionescu, "Most coupled Windows module: Win32k.sys," 2016.

[8] T. Mandt, "Kernel Attacks through User- Mode Callbacks," in *BlackHat*, 2011. [Online]. Available: http://media.blackhat.com/bh-us-11/Mandt/BH{_}US{_}11{_}Mandt{_}win32k{_}WP.pdf

[9] P. Young, "Most coupled Z/System:JES2," 2016.

[10] C. Rikansrud, "Most coupled Z/System:EBCDIC," 2016.

[11] IBM, "IBM ASCII to EBCDIC conversion." [Online]. Available: https://www-03.ibm.com/systems/z/os/zos/features/unix/bpxa1p03.html

[12] F. McDaid, "(E)CKD VS SAN - The Scottish Mainframe Users' Group," in *SMUG*, 2012. [Online]. Available: http://studylib.net/doc/5712216/-e-ckd-vs-san---the-scottish-mainframe-users--group

[13] J. Baez and J. Erbele, "Categories in control," *Theory and Applications of Categories*, 2015. [Online]. Available: http://www.emis.ams.org/journals/TAC/volumes/30/24/30-24.pdf

[14] S. Aaronson, "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*, Barbados, 2016.

[15] A. Winter, "Quantum mechanics: The usefulness of uselessness," *Nature*, vol. 466, no. 7310, pp. 1053–1054, 8 2010. [Online]. Available: http://www.nature.com/doifinder/10.1038/4661053a

[16] V. Putz and K. Svozil, "Quantum music," *Soft Computing*. [Online]. Available: http://tph.tuwien.ac.at/{~{}}svozil/publ/2015-qmusic.pdf

[17] I. Asimov, "ALL THE TROUBLES OF THE WORLD," in *Nine Tomorrows*, 1959. [Online]. Available: http://www.mcguiremarks.com/uploads/3/9/7/9/39793909/isaac{_}asimov-all{_}the{_}troubles{_}of{_}the{_}world{_}(1).pdf

[18] A. Thompson, "An evolved circuit, intrinsic in silicon, entwined with physics." Springer Berlin Heidelberg, 1997, pp. 390–405. [Online]. Available: http://link.springer.com/10.1007/3-540-63173-9{_}61

[19] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete Problems in AI Safety," 6 2016. [Online]. Available: http://arxiv.org/abs/1606.06565

[20] D. Bilar, "The man-machine integration era," *PeerJ PrePrints*, 2016. [Online]. Available: https://peerj.com/preprints/2402v1/