Killswitch Protocols
Or On Engineering Recursive System Death

By Business Class Wine
aka Eric Alston, Seth Killian, and Garrette David

> Industrial man—a sentient reciprocating engine having a fluctuating output, coupled to an iron wheel revolving with uniform velocity. And then we wonder why this should be the golden age of revolution and mental derangement.
> —Aldous Huxley, *Time Must Have a Stop*

Death poses hard questions. This is nowhere more acute than it is for those with the consciousness to value their own existence, such that we mortals tend to take death hard, to the very point of eponymity. A host of protocols for handling an individual's death have emerged over the course of human history. Whether it be the death of a loved one or a beloved leader or celebrity, finally pausing one individual's journey causes others to pause at least momentarily to recognize, grieve, and hopefully celebrate that which was so intertwined with others. Our communities can too be ephemeral, as ghost towns and bands' final tours both evidence, which mean individuals' deaths are more bound up in those of communities than might be immediately apparent.

These mortal tendencies are present as well in our digital lives and the communities that spring up around them. Saved game files for a lovingly micromanaged RPG team or an exceptionally resilient FPS character are inertial data absent an interface with which to re-experience the digital environments that attracted a player in the first place. Examples of this abound in the modern era's pace of technological change, like owning CDs without a CD player and lovingly recalling album art and track lists as evocative symbols of past moments.

These questions are made more salient in cases where a game environment is massively social and persists for years. As Sarah Friend explores in fascinating detail, the death protocols for digital communities merit consideration, not only because these lived experiences matter in their own right, but also because they shed light on a future where our analog and digital lives are increasingly intertwined.[1] One point that brings this into stark relief is that of the "death decision" where an individual or community decides to pull the plug on a virtual environment where people once congregated, which depends on the volition of the individuals designing, powering, hosting, and maintaining these environments.

"EBR-I-SCRAM Button," Wikimedia Commons. commons.wikimedia.org/wiki/File:EBR-I_-_SCRAM_button.jpg

In the cases of profit-driven online role-playing games, pulling the plug has often been motivated by revenue falling enough below costs. Digital community death acutely characterizes the question: when, can, and should the death of systems be engineered? Some systems carry potential consequences so grave that their design is inextricable from the means by which they can be stopped—nuclear reactors in the United States have what is called a SCRAM button, which immediately stops the fission reaction. The first example of a nuclear killswitch was

[1] Sarah Friend, "Good Death," *Summer of Protocols*, 2023. https://summerofprotocols.com/research/good-death

strikingly human, with a man with an ax standing next to a rod suspended by a rope. If the fission reaction began to spiral out of control, the man would cut the rope, causing a control rod to descend into the reactor and halt the process. Nuclear reactors are a striking example of a modern dilemma: we can engineer complex systems that can spiral out of control with staggering negative consequences. Industrialized society, then, needs an overriding failsafe that can stop the system.

Enter the killswitch: an engineered acknowledgment and accommodation of severe and unforeseen consequences. We are so confident that something wildly awful may happen, we engineer mechanisms to stop the entire system, including systems whose continued processes are ostensibly value-producing!

But value can oscillate quickly into costs and at such a magnitude that even precious systems must have a stop. This poses the question of how to shut the thing down, and who should make the decision to do so. While the options that have predominated up to now involve highly centralized control of the killswitch protocol, this form of engineered system death is also the least interesting from a systems engineering perspective because it boils down to designing a killswitch that a concentrated authority can exercise. Put differently, the centralized exercise of a killswitch punts on the hardest questions killswitches pose—who should exercise a failsafe and under what circumstances. While broader input into the exercise of a killswitch may achieve egalitarian motives, this also carries a heightened risk of attack or capture. These tradeoffs between governance options mirror those in most other contexts, although the exercise of a killswitch carries unique considerations that we explore throughout.

Beneath these design considerations is the fundamental question of when intentional death of systems is desirable. Some systems, such as a derelict MMORPG, may have simply run their natural course. One common outcome is that without maintenance the MMO environment can be overwhelmed by pests like gold farmers or cheating players. For other systems, a killswitch is a temporary safeguard to prevent certain system processes from exceeding an intended function, like a circuit breaker that melts to prevent greater damage to the system. In other cases, system death is undesirable, especially if those who control the killswitch exercise it for adversarial reasons. This range in desirability of system death emphasizes the ubiquity of killswitch protocols and the importance of their design. Both considerations centrally motivate our analysis here.

Killswitch Design and Governance

Given how common killswitches are in human engineered systems, including digital communities, protocols have naturally emerged to govern their exercise. A killswitch should not be lightly engaged and, in some instances, cannot be reversed once the button has been pushed. The necessarily overriding power of these protocols within complex systems is clearly a double-edged sword.

At the one end of the continuum of governance choices for killswitches is the total reification of the protocol into an automated set of triggers. While this removes such an override from the control of a central authority, it also removes it from the control of those within the system that is being stopped, overridden, or indeed, killed outright. A killswitch that can only be exercised by a chief executive or similar governing council preserves human discretion, but similarly subjects this to the benefits and costs of concentrated authority. This governance can be checked and balanced as well, with certain systemic overrides or vetoes exercised by an independent authority or by democratic will. This shows how system overrides can vary from pure automation to democratic referenda, depending on the system.

Except for systems for which automated execution is more tractable, like electricity transmission through a series of circuits or the exchange of unitized abstract financial instruments, systems that have a failsafe tend to require human judgment to execute it. As protocols govern complex human systems, they have consequences that people naturally seek to channel or avoid, depending on individual costs and benefits. If unchecked in their authority, those who control protocols can get away with a lot, which means protocols can themselves be dangerous, a concept explored in detail by Nadia Asparouhova.[2]

Midjourney

The risk of abuse of authority alone creates an internal demand for failsafes, in that their existence can check the excesses that unconstrained exercise of power can entail. Such failsafes can shape and constrain the incentives of all system participants, but especially the most powerful if they are subject to killswitch execution by a distinct class of participants in the system. The power of a killswitch thus constrains the powerful. However, within complex systems, there are no panaceas, only solutions that pose tradeoffs. These same constraints tend to come with caps on the total value capture possible within such a system or present a risk surrounding undesirable exercise of the killswitch protocol. Similarly, killswitch automation is not itself a foolproof solution and is impossible given the nature of the system being governed in some instances.

The presence of effective killswitches can therefore directly affect the stance that those governed adopt toward protocols, potentially raising the awareness that there is a more latent system in place and provides an endpoint from which to consider the systemic structure created by protocol design choices. But setting aside total ignorance of the control systems to which we are subject leaves systems that either extract or apportion the efforts of the individuals within them.

These verbs are deliberately chosen to evoke the dystopian or productive character of the collective systems to which we belong, closely mirroring Angela Walch's characterization of protocols' effect: are participants reluctant and therefore resistant or fully conscious and thus willing?[3] The latter class of system participants tends to be more productive. By the same token, however, being "fully conscious" of a protocol is, in some instances, antithetical to what a protocol does. Routinizing human behavior into a protocol can be characterized as a form of forgetting in the same way political engagement can feel exhaustingly endless. Full awareness of protocol engagement seems to reduce or even invert the value of following a protocol in many cases. It's an expensive path that should not be undertaken for all protocols in all places. But a killswitch's existence (and resultant responsibility borne by the stakeholders of the system being killed) may itself encourage protocol awareness.

Checks on the power of the governors tends to induce more representative governance and, therefore, more willing and conscious participation. The mere presence of killswitches can prove a constraint on the excesses of governance authority. Executives subject to reelection or popular recall are forced to countenance the needs of those they govern in ways that dictators never do. Whether the killswitch alone can act against existing authorities or the system itself is constrained more generally, killswitch protocols are a coordination exercise. This

[2] Nadia Asparouhova, "Dangerous Protocols," *Summer of Protocols*, 2023. summerofprotocols.com/research/dangerous-protocols

[3] Angela Walch, "The Protocol System Experience," *Summer of Protocols*, 2023. summerofprotocols.com/research/the-protocol-system-experience

representativity necessarily comes at the cost of expediency. A system without a killswitch is one where executive authority is comparatively unconstrained and can act without the approval of an additional class of system participants. Put more directly, such unconstrained systems can do more stuff: more "moves" are possible in this kind of system than in slower, approval-based systems. In cases of high stakes and high time-sensitivity (e.g., a building on fire, or a global pandemic response), autocratic systems carry an explicit advantage. Thus, a single authority exercising a killswitch protocol is more expedient than a system where the execution of the killswitch requires the input of many parties. The representativity of coordination comes at the cost of efficiency, at least with respect to the cost of reaching the decision itself.

If the only cost of representative killswitch governance was efficiency, design of killswitch protocols would be far easier than it actually is. But in addition to coordination costs, these killswitches carry a far greater risk: distributed input on killswitch protocols increases the vulnerability to adversarial exercise of the killswitch. For example, nuclear launch codes can never be legitimately exercised by someone with adversarial intent—subversion or coercion is required.

In contrast, with publicly traded securities, those with intentions adversarial to those of the existing managers of a firm can gain control of the firm. Opening access to the tools of control is more egalitarian, but this unfettered ability to access the governance rights of the system at a market-determined stock price carries the consequences of transfer of governance rights to third parties regardless of their underlying intent. For those who agree with the motives of recent activist shareholder movements, this type of "democratized" control is normatively preferable.

Democratizing input to governance of killswitches may be understood as closer to a strategy (as opposed to a binary outcome), which can be employed by wide ranges of belief-holders. Indeed, Rafael Fernández explores the limitations and dangers of evoking a swarm, for perfectly distributed control defies centralized command. Swarms' focus and direction are command capture-resistant.[4]

For those who design and engage with complex systems with imperfect control, this tradeoff is a structural constraint. The relationship of increasing distribution and increasing adversarial surface area is not a flaw to be solved but a structural relationship. Like the top and bottom part of a drawn bow, these values move together. Considering them independently may push a protocol towards an unattainable target.

While there are clear costs to democratizing decision-making, there are also substantial risks. But there are still clear net benefits in many instances, as the comparative outperformance of representative and constitutionally constrained governments around the world indicates. Given sufficiently stable periods in which constitutionally constrained systems can more effectively coordinate competing social groups, these systems tend to generate more social flourishing than autocratic alternatives. One explanation is that a system that enjoys legitimacy will motivate its members to give more to the system than a simple rational-interest calculus might suggest precisely because legitimacy fills in the gaps that the uncertainty creates. In private systems where exit costs are sufficiently low, oppressive governance cannot survive in the face of competing systems that offer participants more representative alternatives. There are, of course, other contexts in which centralized governance of killswitches makes sense, which can be tied to massive consequences from exercise (such as using nuclear devices), highly time-constrained environments, or requirement of specialized knowledge to adjudge when exercise of the killswitch is

[4] Rafael Fernández, "Welcome to the Swarm," *Summer of Protocols*, 2023. summerofprotocols.com/research/the-swarm-and-the-formation

appropriate (such as very technical hardware or software systems).

The extent to which a given human system needs to cultivate legitimacy ultimately depends on the purpose that system is designed to achieve and the broader context—including competitors and regulatory environment—in which the system operates. These human-designed and collectively populated systems are most commonly expressed as constituted organizations and therefore a killswitch's legitimizing role within such organizations is greatly determined by the extent to which competing organizational alternatives exist. The intent of a human-designed system cannot be realized absent the context of a protocolized purpose.

Killswitches Straight Killing It

The recursive override function of killswitches that we have described is more ubiquitous than the term might suggest. Our systems frequently require the presence of a method to prevent internal processes from spiraling out of control, to align incentives between system participants, to create a coordinating equilibrium among independent players, or to obtain a more democratic means of governance. As already emphasized, choice in killswitch design and execution carries complex tradeoffs for those subject to a system and its protocols. In this section, we consider killswitch protocols within commercial firms, data trusts, financial markets, gaming tournaments, and DAOs to derive generalizable design principles for killswitches.

Workers and Their Killswitches

A popular finding in the economics literature on organizations surrounds a feature once unique to Toyota assembly lines. In Toyota plants, the employee at each stage of the assembly line had the ability to stop the entire assembly line if they saw an issue or perceived a salient risk. Those familiar with modern assembly line manufacturing well understand how stopping the entire line is tremendously costly in terms of the plant's output and is akin to calling a work stoppage. Furthermore, once stopped, the line can take a substantial amount of time to restart, compounding the costliness of such a decision.

The Toyota killswitch design choice creates a positive feedback loop. When a worker is confident that the urgency is sufficient to stop the line, having that option available to all means the plant is able to get more feedback and in much less prescribed ways. Higher performance resulted from better harnessing the knowledge that was local to each line worker. The decision to distribute killswitch control among all assembly line employees is considered to be a significant input to Toyota's success compared to the U.S. auto industry over the same time period.

While it is likely that this killswitch protocol could only operate in the high trust environment that the company culture created, it nonetheless is a striking signal of the extent to which management trusted its employees to make decisions on behalf of the firm. Trust is a recursive feedback loop and cannot be enforced via contract. Judicious use of the killswitch by workers means Toyota gets high-impact, low-volume feedback from those closest to the product. In a culture of mutual respect, workers feel like they have a voice and are empowered by the trust shown by management.

Photo by Greg Bulla on Unsplash
unsplash.com/photos/white-and-black-switch-on-white-wall-Xlz-BS1BP_Q

Of salient decisions to halt an ongoing process, union work stoppages rank up there. The ability to strike is central to a union's authority to represent workers in a collective bargaining process with a large employer. The centrality of striking to aligning bargaining incentives is such that the right to strike is commonly enshrined in law. Yet the exercise of this particular killswitch is necessarily centralized, as a strike is only effective if a large portion of a given employer's workforce walks out. Indeed, unions can fine members for failing to stop work when a strike is called and are notoriously unfriendly to the "scabs" who cross union picket lines.

Strike killswitches are not limited to labor contexts. The problem for individual workers — who are atomistic and relatively easy to replace compared to a monopsonistic employer — is quite similar to the problem facing users of large internet platforms. An individual user can quit any major platform, but their departure will have a negligible effect on the data the platform gathers, let alone the terms that give them near total control over individual platform users' data.

Just as unions emerged to check imbalances in bargaining power between employers and employees, new data governance models are emerging, complete with killswitches. In the case of the Superset Trust, one of the world's first special purpose data trusts, data contributors' interests are represented by trustees monitoring use, users, and revenue derived from member contributors' data. If the Superset Trust deems data uses or revenue to be at odds with member interests, it can exercise the circuit breaker function and batch revoke member consent to continue collecting or using the data. Like a general strike, the killswitch is not designed with intent for regular use, but its presence outside the control of the data's end users aligns incentives in a way that is structurally analogous to the canonical work stoppages used by unions.

This makes the more general point that systemic similarities can provide fruitful grounds for understanding ideal killswitch protocols for digital contexts, for it is unlikely that a single killswitch protocol is optimal for all digital governance cases.

Overriding Independent Agents

Systems that rely on the aggregated behavior of many independent participants create their own unique challenges. Financial markets depend on a huge number of independent investors whose actions in the aggregate can lead to runaway outcomes like market crashes and financial panics. Tournaments with many individual competitors rely on rules whose intent may not be shared by all participants, a subset of whom would like to win at all costs. Decentralized autonomous organizations (DAOs) frequently aggregate funds for community public goods, but simultaneously create a pot of money that is both an attack vector and an often-undesired killswitch for the DAO itself. As human systems become more complex and more digital, these examples all help to understand the often fraught dynamics of killswitch design and execution.

Midjourney

Capital markets tend to have opening and closing times, before and after which trading is highly constrained relative to the volume that major exchanges have come to display. Yet in addition to well-known periods that capital markets are closed for exchange, these markets tend to reserve the ability to halt trading in certain stocks or even all market activity under sufficiently extraordinary circumstances. Argued as a limitation on market sentiments that can catalyze into a panic, stock exchanges halting trading in a given security is intended to serve as a pressure release valve, enabling cooler heads to prevail during the next period of trading. If a company's indicators or related news event are sufficiently negative, a halt does not prevent markets from ultimately working, for such a stock is likely to suffer once trading reopens, absent a change in information or sentiment.

Financial market interventions display another important institutional characteristic of killswitch protocols: they can be tailored to specific components of a system. Stock exchanges can suspend all trading in extraordinary circumstances; a government can further intervene and close markets. Together these abilities display a characteristic noteworthy of nested killswitches: specific functions or the whole system can be subject to an override, depending on the trigger.

Interventions into a market of many independent participants can also have the opposite of their intended effect, either by creating a buildup of sentiment that is unleashed in a flood when trading reopens or through restraining the ability of liquidity to ameliorate the negative outcomes a shutdown was intended to dampen. Limits to constructivist intent within complex systems means killswitches will in important instances fail to fully restrain the system's processes or carry unintended consequences when they do manage to do so.

Even for a video game tournament (with increasingly large sums of money at stake), the rules tend to include motivations—the why in addition to the how. This is in part to cover potential rule violations that have no defined shape in gameplay itself, such as collusion among players. One common solution entails placing trust in a tournament director or general manager to adjudicate potential rule violations that are not definable in the code that governs individual play. This is necessarily a high trust evaluation—a killswitch for a competitor earning outsized returns on unanticipated strategy that is sufficiently outside the bounds of fair play as broadly understood. This looseness in rules, decided at the sole discretion of tournament directors and general managers, dissuades players from pursuing meta-strategies beyond the object level of the game. Any competition will naturally invite players to pursue value at the margins (e.g., improved seeding, extra time to make a decision, side preference, etc), but without this oversight, the risk of outsized rewards where someone wins by playing around the rules rather than within them increases. Cheat-like behavior is very hard to distinguish from excellence. Worse, the wider the participation in the protocol and the more value stored, the more one can expect these incentives to distort behavior. Discretionary disqualification is the threatened killswitch in these contexts that serves to align players' incentives to contest within the rules, as opposed to skirting them in ways system designers never intended.

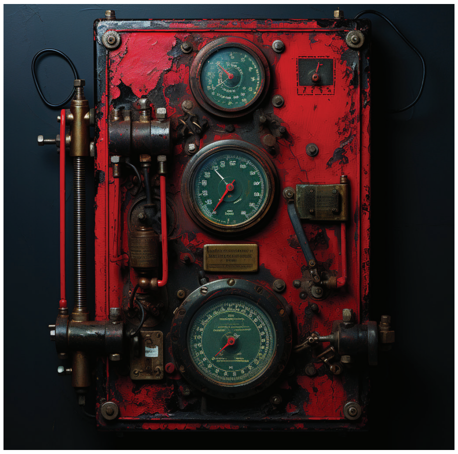
While a retelling of the DAO hack on the Ethereum network would exceed the constraints of this article, the now-infamous episode in cryptocurrency history has several generalizable lessons pertinent to understanding killswitch protocols. A hacker discovered a bug in the way money could be withdrawn from the collective investment vehicle that was "the DAO," and exploited this to siphon off a large amount of ETH. Most obviously, this event shows how exit from a system can itself be an attack vector and cause the death of financialized DAOs, both through unauthorized withdrawal, but also due to the potential collapse of the beating financial heart of a given DAO. The DAO hack was big enough to pose an existential threat to the Ethereum core protocol and effectively taint all funds in the market.

The DAO hack is a specific example of the general phenomenon of how exposure to financialization or other tradeable within-system goods creates attack vectors. These vectors are naturally pursued by self-interested agents within the system to the degree that the original system is successful in capturing value. This pattern is so structural it is likely something that system designers must manage, as opposed to being something that can be "solved" through mechanism design. DAO deaths have not been limited to the original DAO whose spectacular demise led to the birth of Ethereum Classic. Many DAOs need operating funds in order to engage, which has resulted in transaction fees on a particular DAO being partly diverted into a public treasury that can be used to fund public goods for the DAO. As several DAOs like Rook and Mango have found to their chagrin, subjecting disbursement of treasury funds to a known tokenized voting rule legitimizes a raid according to the rules, with groups like “Risk Free Value Raiders” organizing to deplete treasuries by buying governance tokens sufficient to sway funding outcomes at a cost lower than the amounts being disbursed. Risk free value, indeed!

What Can Our Killswitches Teach Us?

At a more structural level, the issue of clearly defining the “public good” for a given system itself outstrips the ability to define all margins of this good. In the case of DAOs, certain components of organizational processes lack sufficient resolution for algorithmic reconciliation, which then leave it up to governance token holders to determine what is in the collective good. While lining the pockets of current token holders is not a typically defensible public good, there is an interesting argument to be made that depleting a common treasury actually removes the fiscal engine that powers a managerial group exercising authority over the DAO commons, such that even raiders may have a narrow public good argument upon which to ground their clearly self-interested actions. What if raiders are merely correcting the system to be truly decentralized and therefore not subject to stringent regulatory authority? While we take a dim view of the nobility of DAO raiders’ motives, this example does emphasize how the justification for exercising a given killswitch can very much tend to be in the eye of the beholder, and relies on the understanding of a given community or system’s purpose that outstrips cleanly automatable triggers like the amount of electricity intended to run through a system.

Killswitch existence and proliferation means that systems designers accept that running a system over time is expected to generate novel and unanticipated costs. Stuff happens. The environment shifts. The incentive structure over time distorts behavior in ways that alter the environment for which the protocol was originally developed. Trust concentrates risk while trustlessness can distribute risk. Yet both high trust and low trust approaches to protocolization can create attack vectors. The risk associated with these attack vectors can be mitigated, but generally cannot be completely removed in a cost-effective, value-capturing way. Either approach will return different value depending on the environment in which it is enacted. A great protocol operating in the wrong context can be as disastrous as a bad protocol. No object-level set of rules can block attack vectors targeting the rules themselves. You can invoke a fuzzier “why” behind the point of the protocol to adjudicate meta decisions (like engaging a killswitch) and this can work admirably (as in the case of the video game tournaments), but will lack the resolution required for automation. Indeed, an early AI told to “win” a racing video game obtained maximum points by capturing valuable side items, but lost spectacularly due to the system designers forgetting to enshrine winning the race itself as a primary objective!



Midjourney

Successful protocols harness value and concentrated value increases risk. Value attracts treasure hunters and rewards treasure-hunting mindsets and strategies, which naturally limit the trust a system can obtain. As any dragon will tell you, accumulating a treasure hoard of value naturally invites treasure hunters, of many motivations. Fruit trees attract sugar-seeking pests. Concentrated fruit trees (as in an orchard) may attract pests in such a concentration that they may collectively—through no specific desire or intention—destroy the orchard and its associated value entirely. This is also a risk associated with private equity corporate takeovers. Importantly, though, value can take many forms, from gold to fiat currencies to social status to familial obligation, such that this is another margin of consideration for killswitch design and execution—what type of value does the system concentrate?

Another distinction between killswitches that our examples illustrate is between killswitch execution in adversarial versus cooperative contexts and how the chosen protocol can partly determine the level of cooperation a system can obtain. In the Toyota example we saw a case where killswitch control created a more productive work environment by facilitating a high-trust environment. In contrast, unionized workers' right to strike tends to be enshrined in collective bargaining agreements themselves, as are other components of the employment environment that employees tend to care about deeply. But rigid delineation of these terms creates a more adversarial stance, with workers (and union bosses) demanding every iota of what they're entitled to and employers interested in ensuring that not one iota more be granted. Killswitches can thus be part of a protocol that cultivates trust among system participants, or one that cements the adversarial nature of relationships within the system. While both high and low trust environments are likely to persist into the future, it is worth considering this specific feature of killswitch protocols when engaging in system design in the first place.

At a system level, our examples show that the ongoing vitality of a system may be inextricable from the death decision. A common pool of money over which to coordinate can become a DAO's fatal flaw just as a common set of rules under which competitors should abide tend to require human judgment for the edge cases where someone wins by subverting the intent of the rules. More generally, these examples also show the risks of decentralized control of system processes, including killswitch protocols, for exit options can be exercised by those with orthogonal or adversarial intent to the resilience and flourishing of any system. Who controls the killswitch also controls the system's survival, to put it most bluntly.

———

The etymology of circuit breakers tracks their first use by Thomas Edison, a use which only grew with the tremendous spread of electrification into the twentieth century. The words fail-safe and (mechanical) override are tied etymologically to the late 1940s, as the widespread uptake of human-engineered systems like airplanes and automobiles increasingly placed their users in places of mortal risk. Killswitch has a less well-defined etymology, although its emergence is clearly tied to the presence of shut-off switches in a variety of increasingly complex and risky machinery. In many instances, these automated shut-offs are triggered by the separation of the human in control of the machine from the machine itself. A speeding motorboat pulling a water skier can rapidly become a death sentence in the predictable moment when a rogue wave knocks the driver out of the boat; many machines can make nightmares of their intended functions—even paper clip manufacturing may become a deadly objective when pursued by fully automated agents.

Yet the need for a recursive override governing complex systems is not limited to the mechanized contexts that gave birth to killswitches. Indeed, political concepts, like using separate government authorities to check one another and popular referenda as an executive recall mechanism, are ancient compared to the invention of the airplane and automobile. Public governments and private organizations are complex systems with potentially perverse consequences if left unchecked. Due process has come to mean a substantive check on procedurally sufficient processes. Most recursively, even as fundamental of a ruleset as a constitution must contemplate its own means of amendment. Thus, the beneficial effects that institutional design provides in coordinating collective action in complex social orders carries its own endogenous need for killswitches.

In the cool new digital frontier involving transparent and distributed governance of data and coordinated units of account, we're seeing increasing automation of components of organizations, as well as governance contexts that facilitate greater distributed control of system components, including killswitches. Vexing recursion abounds here, though! Increased automation begets an increased need for killswitches both to ensure this distributed control, but also concomitantly greater dangers, per the increased costs and risks we have identified. One cause for hope surrounds how software tends to depend integrally on hardware, such that software that cannot be killed through its own code may still have its own Achilles' heel when it comes to pulling the proverbial plug.

Killswitches' emergence within complex systems suggests an inherent efficiency to their designed presence. Control of a killswitch can vary from fully automated to highly distributed and any choice along this continuum of killswitch governance should be assessed considering the context in which the killswitch will be executed. Automation is predictable and not subject to human subjectivity once implemented, but is also rigid and thus lossy relative to the dynamic demands of an unpredictable world. Centralized control benefits from specialization and speed, but can result in capture or misalignment. Distributed control tends to be more representative and transparent, but also is more costly and subject to special interest influence. More acutely, killswitches create a unique attack vector for adversarial interests that can obtain sufficient influence within a distributed system, making the question of killswitches (and their close conceptual bedfellows) a critical consideration for protocol designers of distributed networks, both because of their relative benefits in semi-automated coordination contexts, but also because of their unique benefits and risks in furtherance of the dynamic representativity for which many of these systems strive.

Killswitch protocols are an increasingly essential design component within complex human-engineered systems. They are also eponymously deadly and functionally so by design. Their increasing prevalence in semi-automated digital organizations means greater attention should be paid to their history, design, and inevitable shortcomings. It is our hope that our brief survey introduces their relevance to the protocol designers confronting this brave new world. Δ

ERIC ALSTON is a Scholar in Residence in the Finance Division at University of Colorado Boulder. Eric's research is grounded in the fields of institutional and organizational analysis & law and economics, and explores constitutions, economic rights on frontiers, and digital governance specifically. Eric is also currently engaged in governance design for several distributed network projects.
www.colorado.edu/business/leeds-directory/faculty/eric-c-alston

SETH KILLIAN is an American game designer best known for his work on competitive games such as Street Fighter and Fortnite. He is a cofounder of the Evo Championship Series, one of largest and longest-running gaming competitions in the world. The size and dedication of gaming audiences offer practical insights into the function of cooperative and competitive protocols (from hardcoded win conditions to more nebulous social expectations between players) at scale, over time, and under duress. www.linkedin.com/in/seth-killian-8858ba3/

GARRETTE DAVID Interested in crypto market physics on a standalone and macro basis. Seven years in cryptoland across mining, capital markets, protocol launches and degeneracy. Former neuroscientist.

ProtocolKit

RETROSPECTUS

NEWSLETTER

ProtocolKit
summerofprotocols.com
hello@summerofprotocols.com

ISBN-13: 978-1-962872-18-8
Printed in the United States of America
Printing history: July 2024

1 3 5 7 9 10 8 6 4 2

ISBN 978-1-962872-18-8
9 781962 872188
90000>