Feature Commentary: Organizing for Innovation: Lessons from Digital Counterterrorism

By Brian Fishman

Digital platforms were slow to build robust teams to counter threat actors, but today, many of those corporate teams have robust processes, specialized tools, and innovative approaches to countering highly adaptive adversaries. They operate in a tremendously dynamic environment where their adversaries can innovate at low cost, primarily because of the nature of the digital "terrain" where the conflict occurs. And while the actions these teams take are not kinetic, as those sometimes utilized in geopolitical conflict, the cat-and-mouse game between Trust & Safety teams and threat actors online suggests lessons that are increasingly relevant to the national security community. This article explores five factors that were key to facilitating innovation in Facebook's approach to countering the Islamic State—and that I argue are more generalizable. They are: people, organization, legitimacy, tools, and collaboration. It also identifies lessons that can be learned from that experience. For example, we did not prioritize using a particular technology or focus experimentation in some bespoke "innovation center." Rather, we succeeded because we were made responsible for a critical mission, were unencumbered by past process, and had the right team structured to reduce external dependencies for innovation. Basic technological innovation can occur in an ivory tower, but applied innovation requires proximity to real-world missions. You cannot expect dramatic innovation without failure and iteration in an environment of real responsibility. Fundamentally, that means that innovation requires accepting risk. The structures and incentives of Silicon Valley cannot and should not simply be grafted on to our national security infrastructure. The rewards and costs of failure are different. But military organizations should shoulder the risks associated with innovation and study the lessons of corollary efforts in Silicon Valley and the private sector more broadly.

Over the past 30 years, technology companies built the modern internet—and with it a slew of new methods for communication and commerce. In doing so, they also inadvertently constructed new digital terrain for threat actors to exploit. In order to safeguard the communities and commerce that emerged online, and under significant pressure from governments and civil society, these companies belatedly built mechanisms to identify, disrupt, and deter those threat actors. Collectively, those activities are a key element of what professionals call Trust & Safety.[a] Trust & Safety is a practice of adversarial adaptation mediated by technology that often results in punitive action. And while the actions taken by Trust & Safety teams are not kinetic, the technology, organization, and centrality of technological adaptation necessary for Trust & Safety offers lessons for military leaders now and in the future.

The fundamentally adversarial nature of Trust & Safety drives innovation by attackers and defenders. When I arrived to lead Facebook's efforts against the Islamic State in the spring of 2016, the prevailing instinct among engineers was to build AI-driven classifiers to find content supporting the group. But I understood how the Islamic State's propaganda operation functioned, both on and off Facebook. There was a more straightforward, intelligence-driven way to disrupt the group's formal propaganda operation, which was our initial goal. So, we used vendors to collect emerging Islamic State propaganda on Telegram; established pipelines to triage, label, and hash it quickly; and then were able to detect that propaganda as soon as it was uploaded to a Facebook server.[b] I asked for entirely new ways to measure operational success—built around time rather than scale—and eventually, we regularly ran that process more quickly than Islamic State supporters could upload the first instance of a piece of propaganda to Facebook.

This was a good, creative win, but it was also only a single blow in a much longer cat-and-mouse game. Predictably, the Islamic State

a The Digital Trust & Safety Partnership defines Trust & Safety, broadly, as: "The field and practices employed by digital services to manage content- and conduct related risks to users and others, mitigate online or other forms of technology facilitated abuse, advocate for user rights, and protect brand safety." "Trust & Safety Glossary of Terms," Digital Trust & Safety Partnership, July 2023.

b A perceptual hash is a method to convert a file into a series of numbers. This digital fingerprint can then be algorithmically compared to other such fingerprints to identify media that is similar. Hash-matching is a common method to identify to child sexual abuse material (CSAM), terrorist propaganda, and non-consensual intimate imagery (NCII).

Brian Fishman is a co-founder of Cinder, which builds an orchestration platform for Responsible AI and Trust & Safety. He previously led Facebook's work to counter terrorism, hate groups, and large-scale criminal organizations. Prior to Facebook, Fishman served as the director of research at the Combating Terrorism Center at West Point; ran Palantir Technologies' disaster relief program; worked in-house at New America, a think-tank; and worked on Capitol Hill. Fishman authored The Master Plan: ISIS: al-Qaeda, and the Jihadi Strategy for Final Victory *(Yale University Press), which was built on a pathbreaking course he taught at West Point in 2008 about the Islamic State's plan to govern.*

innovated: by speeding up their process, editing core material to confound detection tools, and eventually operating on Facebook in more informal ways. The lesson is neither that AI classifiers are too clunky (they are, in fact, very useful) nor that lower-tech solutions produce partial victories. Rather, it is that technology must fit the mission and that every victory is fleeting against innovative opponents, especially online where the cost of iterating is low.

So, how do you build systems able to innovate and integrate technology into complex, adversarial missions so that you can strike not just one blow but an entire campaign? In my experience, five factors stand out: people, organization, legitimacy, tools, and collaboration. In short, successful innovation requires the right people, which are sometimes atypical for your organization; the right organizational structures and disposition; mechanisms and leadership to establish and maintain the legitimacy of new processes; technical tools designed for flexibility, innovation, and impact (not point solutions or flashy demos); and a commitment to use technology to improve collaboration across organizations and sectors.

Before exploring those five factors in depth, this article briefly describes the history of Trust & Safety and notes unique features of this sort of digital contestation. The purpose of the article is to suggest mechanisms to enable technological innovation, but, perhaps counterintuitively, most of the recommendations regard traditional issues of personnel, organization, and leadership. That is because applied innovation is more a matter of adapting workflows to capitalize on emergent technology more than it is advancing raw science or operating on the bleeding edge of what can be achieved with physics or biology. Applied innovation requires openness to cutting edge technology, but fundamentally, it is about matching technology and organization to the mission—and preventing legacy processes from getting in the way.

Background & Key Concepts

Trust & Safety has a longer history than generally understood and some key features that shape how the competition between threat actors and Trust & Safety professionals plays out.

History of Trust & Safety

Trust & Safety efforts began in earnest in the late 1990s when companies such as eBay organized to counter fraud, counterfeits, and other disruptions to their digital marketplace.[1] Social media companies like Facebook and YouTube were slow to develop robust Trust & Safety teams, but have since built some of the most sophisticated operations for building and implementing private policy anywhere in the world.[c]

At the significant risk of oversimplification, Trust & Safety practices can be bucketed into two intertwined categories: community management and threat disruption. Community management governs how people engage each other on a platform, so the rules vary from one site to another. For example, a platform built for discourse might allow more contentious political speech or sexualized content. Conversely, a site for buying and selling vintage T-shirts might decide it is not the place for such content. In both cases, community management generally requires delineating rules, communicating them to users, and aiming to correct bad behavior before taking irrevocable enforcement actions.

"Platforms shape the digital terrain itself, not just the countermeasures they use against threat actors. This is a powerful, but limited, advantage."

Threat disruption is different. It is focused on identifying and disrupting actors engaged in deeply problematic behavior, sometimes offline: terrorists, child predators, organized criminal networks, and nation-states. Most platforms have policies that prohibit these actors, but many lack the resources to enforce them aggressively, which requires defining, identifying, investigating, acting against, and then defending against their shifting tactics. These actors are often the worst of the worst, but they are also less common. So threat disruption requires finding needles and networks in immense haystacks of data.

Scale, Terrain, Account Regeneration, and the Villain Use Case

The conflict between threat actors and Trust & Safety professionals has some unique features. The first is scale. A large-scale Trust & Safety operation makes millions of decisions daily about individual pieces of content and accounts. In Q2 2024, Meta removed 7.5 million pieces of content just for violating its rules around terrorism.[2] This means that both human and automated systems must be built to process very large amounts of information and that even a low error rate, whether false positives or false negatives, can result in a large number of bad decisions. In a highly scrutinized space, those errors can draw regulatory pressure and alienate users.

It is tempting to conclude that the scale and sensitivity of these choices creates a simple operational tradeoff: the scale of these decisions requires automation, but their sensitivity demands the thoughtfulness of human decisions. That tradeoff does exist, but the basic version is over-simplified. The reality is that human decision-making at scale is extremely error-prone. Even before the current explosion of AI tools, AI systems at Facebook (and other methods of automation) were regularly as accurate as human beings at many Trust & Safety tasks. But they could also be expensive to train and made mistakes that were stranger and more inexplicable than those made by humans. It is not just the scale of the mistakes AI and automation can make; it is the nature of those mistakes that can make them more problematic, even unacceptable. Nonetheless, it is important not to assume that humans do all jobs more accurately (in aggregate) than AI and automation more generally.

The second feature of the conflict between threat actors and Trust & Safety professionals is that the platforms shape the digital terrain itself, not just the countermeasures they use against threat actors. This is a powerful, but limited, advantage. Platforms build the algorithms that surface content, determine how easy it is to find new accounts to engage, and decide how much privacy to build into a digital system. Trust & Safety teams often advise on these systems to highlight potential risks. But just as the walls of a medieval city might be constructed both for security and to enable

c While this discussion primarily draws on lessons from the largest and most sophisticated Trust & Safety efforts, it is important to note that smaller teams face related challenges and sometimes innovate extremely effectively as a result.

everyday life and commerce, digital platforms are not constructed solely, or even primarily, to hinder the behavior of threat actors. Platforms are often designed to allow likeminded people to find one another; buyers to find sellers; and a range of users to engage with various levels of privacy and broadcast functions. The ability to shape this terrain gives platforms a huge advantage—both in terms of designing for safety and in gathering intelligence. But that advantage is not fully utilized, even by well-meaning platforms, because the same systems exploited by threat actors are also used by productive users—creating both a commercial tension for platforms and one of general social utility.

The third feature of the threat actor versus Trust & Safety contest is that threat actors can regenerate capacity online, often at minimal cost. This means that threat actors can iterate and experiment tactically and operationally at a scale that is simply not feasible offline. If their accounts are removed, they can recreate them. If a particular content type is discovered, they can move to another. Sophisticated platforms can make this innovation less fruitful, but they cannot eliminate the process. Viewed from the perspective of Trust & Safety, the physical world can represent a safe haven for digital threat actors, even when their ultimate aim is harm in the real world.

The internet beyond the 'walls' of a particular platform also serves as a safe haven. Cross-platform collaboration against threat actors remains nascent. When Facebook disrupted Islamic State operations, its supporters could (and did) plot and plan on Telegram to circumvent those techniques. There are some cross-platform coordination efforts—addressing child sexual abuse material (CSAM), non-consensual intimate imagery (NCII), terrorism, and disinformation—but they are not systematic enough. At the same time, a platform's only ability to impact a threat actor in the offline world is to inform relevant law enforcement authorities. There are very impactful examples of this kind of collaboration working, but such mechanisms are limited given the global nature of the internet, law enforcement capacity, and the unreliability of law enforcement in some jurisdictions.

Finally, every digital tool is dual-use, even those developed to mitigate harm. Product managers sometimes imagine a 'hero use case,' which essentially reflects an ideal user that fully embraces a product to get the most out of its functions. But for every hero use case, there is a villain use case, whereby some actors use the same tool for harm. For example, early in my tenure at Facebook, user reports of terrorist material on the platform were erroneous more than 90 percent of the time. Some of these reports represented overzealous users with generally good intent, but others were deliberately reporting benign content as terrorism in the hope that Facebook would be more likely to remove it. Every technical system that creates capability also creates new attack surfaces.

People, Organization, Legitimacy, Tools, and Collaboration

There is no silver bullet to create innovative systems. But the five factors of people, legitimacy, organization, tools, and collaboration are critical.

People

The mission of Trust & Safety teams is ultimately to make a platform safe and thereby inviting for the majority of users. In that respect, it is deeply aligned with the commercial mission of most technology companies. But the process of highlighting risks, expelling some users, and embracing paranoia as a professional virtue is non-standard in generally optimistic Silicon Valley. Unsurprisingly, Trust & Safety attracts a mélange of professionals somewhat different than the Silicon Valley workforce as a whole—and one that is more focused on the risks of a platform rather than the potential benefits to the wider community.

"Highly process-driven organizations often resist innovation because individuals in them are rewarded for implementing that process rather than achieving mission-altering outcomes."

It is treacherous to synthesize complex personalities into typologies. Nonetheless, I like to think about three basic "personas" in Trust & Safety: 'tech do-gooders,' 'the ones who know,' and 'hunters.' Tech do-gooders believe in the general social value of technology and that to realize those benefits the risks and costs associated with technology must be mitigated. These folks often have engineering, product, or design skills and would have a place in tech companies even if they were not working on Trust & Safety. The-ones-who-know have seen first-hand the downsides of technological innovation. They often come from marginalized communities at-risk online and have linguistic, cultural, and lived experiences far more diverse than technology companies writ large. For example, Trust & Safety as a discipline has more women in leadership roles than tech generally, and Trust & Safety includes incredibly diverse groups of people that speak the languages and understand the cultures of global communities. Finally, there are the hunters. These are folks who relish the fight against bad actors. They often think of themselves as protectors. Many now come from law enforcement and intelligence communities and sought roles at tech companies because technology is now a key terrain for the threat actors they pursued elsewhere. Yet, the tech community has grown some of these people internally; they often grew up fighting spam and fraud.

All three of these personas are necessary for Trust & Safety to succeed. The tech do-gooders often understand technology best and can imagine ways to utilize cutting edge tools. The-ones-who-know understand how those new techniques will work and can apply them in various contexts. Although Trust & Safety tends to embrace diversity, these people are often the most junior members of a team. Nonetheless, they are often where the rubber meets the road and regularly are sources of the on-the-ground knowledge that is necessary to keep pace with adapting adversaries. Finally, the hunters have the experience and skillset to target the worst-of-the-worst actors. They think in terms of networks, organization, and the nodes that have an outsized impact. For innovation to work in an adversarial setting, all three personas are necessary, and that means that technology companies have to recruit people that do not fit their standard profile.

Organization

Highly process-driven organizations often resist innovation because individuals in them are rewarded for implementing that process rather than achieving mission-altering outcomes.

To incentivize innovation, organizations should limit process, reconsider personnel assessment, and embrace experimentation, despite the reality that it will inevitably lead to some failure. Crisis often enables such structures, but they can be implemented without crisis by leaders willing to accept the risks.

Many of these factors were present when I joined Facebook, and they contributed to an environment I was able to utilize effectively. The Islamic State was (belatedly) seen as a true crisis; we had a cross-functional team whose participants were unusually independent of their 'home' bureaucracies; and resources were plentiful. Finally, we had leadership clarity, meaning both that Facebook's most-senior executives supported the work and that I, as the operational leader—a relative outsider with subject matter expertise and the credentials to prove it (not the same things; both important)—had unusual credibility and leverage to try new things.

Innovation in conflict is difficult because the importance of the mission can lead to an ethos where failure is inconceivable and unacceptable. That notion is sometimes necessary, particularly at a tactical level. But failure and iteration are critical to applied technological innovation. Organizations, and the leaders that guide them, must facilitate experimentation and celebrate productive failure. If not, they will disincentivize the risk-taking that is necessary for new ideas, technologies, and procedures to emerge.

When I arrived at Facebook, the community management elements of the Trust & Safety effort were generally divided into three major bureaucratic components: policy, operations, and engineering. These teams worked together, but individuals within those verticals were accountable to their own leadership. Leaders of those teams sought unity, but that intent could break down because distinct organizational perspectives were codified not just in mission prioritization from leadership but in bespoke personnel assessment standards which were not turned primarily to the success of the cross-functional group.

The Dangerous Organizations and Individuals (DOI) team that built a renewed campaign against the Islamic State operated differently. For starters, it was extremely well-resourced, more than 300 people strong. Moreover, the DOI operations team had its own technical capacity—data scientists and engineers who could explore new ideas quickly and without cross-functional handwringing. Finally, the engineers seconded to work with this DOI cross-functional group were also 'graded' (especially early in my tenure) by their own organization based on the importance of the work rather than compared on narrow metrics, which was more standard within the engineering organization.

It was ultimately valuable to have technical capacity both embedded in the operations team and engineers seconded from the engineering team. The former allowed us to iterate quickly and test new ideas with minimal friction; the latter emphasized scalable process and quantitative success metrics. Notably, the traditional engineering teams were paid more and generally 'better' engineers. Their processes and products were generally more rigorous. But in an innovative, adversarial environment workable is better than perfect—and so the technical creativity of the operations engineers pointed the way toward solutions that could subsequently be scaled.

Legitimacy

Leadership is critical in an organization innovating with technology in an adversarial environment. Process-derived legitimacy is too slow and outcomes can take time, particularly when the adversaries adjust. Leadership is therefore critical, both at the strategic and operational level. The strategic leader must generate resources and space to break standard procedures, including over prosaic issues such as personnel assessment; and tolerate missteps and imperfection. The operational leader must generate clear priorities; insulate the operational team from inevitable bureaucratic politics; and ensure that operational wins can be translated into strategic ones. The art of the innovative operational leader is that they must direct the team when necessary and enable innovation to bubble up organically.

"Leadership is critical in an organization innovating with technology in an adversarial environment. Process-derived legitimacy is too slow and outcomes can take time, particularly when the adversaries adjust. Leadership is therefore critical, both at the strategic and operational level."

Sheryl Sandberg, then the chief operating officer of Facebook, created the strategic space for Facebook's campaign against the Islamic State, and I was the operational leader tasked with designing and executing it. Fairly or unfairly, my legitimacy as a *credentialed* expert on the Islamic State was critical. Before my arrival, Facebook already had analysts that understood the Islamic State; it had relevant linguistic and cultural expertise rivaling any intelligence agency; and it had tremendous engineers with more data than they knew what to do with. But my knowledge of the group coupled with credentials, ability to communicate at a senior leadership level, and willingness to accept personal responsibility and risk for new techniques was key to unlocking that latent capability.

Coalescing the cross-functional team to execute those plans was primarily my responsibility, but managing the complex bureaucracy of a major corporation is no small task. This only worked because my leadership coached me on how to engage Facebook's top-level decision makers. Moreover, they avoided the mistake of many leaders in a crisis-driven organization, which is to reward folks for reacting well to crises, but failing to reward people for preventing crises in the first place.

This set-up worked. In just over a year, Facebook went from finding almost zero Islamic State material proactively to identifying 99 percent of the terrorist material it removed via automated systems.[3]

Legitimacy is critical for generating innovation, but maintaining that legitimacy is more difficult than it appears. The reason is that innovation fundamentally requires failure. This ethos is built into the bones of Silicon Valley, where the "power law" of venture capital stipulates that most financial returns will be concentrated in a small percentage of startups. Others will break even, and many will fail completely. The "power law" means that even the people supposedly best at identifying innovative concepts and teams recognize that they will fail most of the time. They still win big because a single

major success can outweigh numerous small failures. Such a pattern is not easily applicable to military affairs or geopolitical issues more generally. It is rare that occasional big victories compensate for repeated failures. Nonetheless, innovative military organizations must allow for mission-relevant experimentation if they are to produce a culture that enables groundbreaking ideas and innovation.

This will be extremely difficult to achieve. For strategic leaders, it will mean carefully selecting missions where higher-risk, higher-reward approaches can be tested. It also means adjusting communication patterns to prepare stakeholders for risk. Innovative operational leaders must communicate clearly with superiors about risks, and those superiors must not only accept, but champion, them. Combatant commands must communicate up the chain and political leaders in the executive and legislative branches ultimately need to bless experimentation. Publicizing experimentation is important as well. Failure costs money, time, and in some awful cases, lives. But failure is not always a scandal—if the risks are well-considered, the mission critical, and innovation necessary. Innovators should engage the media and related stakeholders early, educate them on the risks, and explain that adversarial shifts demand creative approaches that will inevitably be imperfect, especially initially.

Tools

The most visible manifestations of innovation are not necessarily the most important. Over and over again at Facebook, we identified internal tools that failed to provide accurate information, conflicted with other tools, or were built for static challenges, not dynamic ones. Innovation requires fast iteration and adaptation, and that means building core tooling capabilities that enable operational and tactical creativity. Innovation means expecting obsolescence from technologies and processes—so you should emphasize core technical platforms that are easily updated, extensible to a wide range of other technologies, and modular enough to facilitate process and technological dynamism.

In 2016, Facebook had some dynamic systems but not others. For example, Facebook had incredibly powerful tools to query immense datasets and map entities related to one another. These systems were relatively easy to use and accessible to many people in the company. That meant that frontline data scientists could query information and test hypotheses almost as quickly as I could generate them, which allowed us to quickly identify promising concepts to disrupt Islamic State activities. At the same, Facebook did not have good tools to visualize networks, enable non-technical subject matter experts to reliably fanout through them, or quickly construct new enforcement procedures. In some cases, it could be difficult to understand how or why a particular enforcement action had been taken - in part because there were multiple, sometimes conflicting systems for gathering that information. We had very powerful AI systems, but they took too long to retrain and deploy.

That meant that we could not always update actual enforcement systems as quickly as the Islamic State could adjust—and when we did, it was often by updating human-driven processes as opposed to technical ones, so we did not systematically capture data on their adversarial responses to our improved process. Those data limitations might have been damning in Facebook's traditionally metrics-driven decision processes, but the unique organizational and leadership structure of the DOI XFN meant that during key time periods we could adapt regardless.

Nonetheless, that was a poor substitute for having better, more dynamic systems to begin with. Improved basic tooling was critical to long-term innovation. Large bureaucracies cannot scale innovation forever based on the credibility of individual leaders. So, Facebook invested. Better mapping software powered network-level takedowns of terrorist material. Improved AI training meant classifiers could better keep up with current trends. Consolidating competing tools that sometimes produced divergent information reduced confusion and ensuing decision slowdowns.

Notably, most of this innovation was focused on capturing and understanding signals, rather than innovating the sort of actions we took against the Islamic State. Improving our own decision-making was more important than improving the precise actions we took against threat actors. (It is worth noting that other teams did innovate more in the actions they took against other threat actors, but this was less impactful in the DOI context.) The success was primarily tooling and innovation built to derive understanding from data, to drive decision-making, and to build components of operational systems that could be easily rearranged in response to changing operational and tactical demands.

Collaboration

Like other harmful actors that operate online, the Islamic State does not simply use one platform. It might coordinate internally on Rocket.chat, advertise propaganda on Telegram, recruit on Facebook, and store content on Dropbox. A single digital operation might span five or six platforms. As a result, improving Facebook's defenses has had a limited impact on the group as a whole and left key elements of its digital network intact. This means that, as in traditional geopolitical competition, coalitions are a key part of confronting harm online. These collaborative spaces are also a venue for technical innovation, but they pose unique challenges.

First, innovation is a full-time job. Time-bound efforts deployed in a 'hackathon'-style environment might generate new ideas, but they are unlikely to produce products that can be used over time. It is possible to build joint organizations with generic mandates to innovate, but the distance of such bodies from tactical realities will limit their understanding of the adversarial environment and reduce their urgency to innovate. Innovative joint (and combined) organizations must maintain staffing for an extended period. Seconded personnel should access tactical leaders from their home organizations to generate ideas and vet progress, but if those seconded personnel are not exempted from the typical personnel reviews of their home institutions, they will likely be less innovative.

Second, some coalition partners will represent best practices in any coalition and will likely have existing tools that can be appropriated for new purposes. At Facebook, I helped build the Global Internet Forum to Counter Terrorism (GIFCT), a coalition of tech companies dedicated to sharing tools and processes to counter terrorist activity online. One of GIFCT's core tools is a database of hashed terrorist propaganda. Participating companies upload hashes of terrorist material so that others may download them to identify that material on their own platform. This basic idea was originally used to counter child sexual abuse material (CSAM) and the technical platform used for GIFCT hash-sharing was originally built to share hashes of malware. But an enterprising engineer at Facebook recognized we could repurpose that tool (called Threat Exchange), and I was able to convince internal stakeholders

and other companies to use it for a new purpose. Sometimes, technological innovation is simply recognizing that an existing tool can be used for a different mission. This may not energize engineers and those excited by using cutting edge technology, but this is particularly useful when the mission is elevating the baseline capability of a coalition.

Third, building innovative shared resources does not mean that coalition partners will use them. Facebook had the resources to integrate its internal tools for detecting media hashes to the GIFCT database. Facebook could both push and pull those hashes seamlessly. But many smaller companies did not have the resources to integrate with the shared database nor, perhaps, even the ability to store and match hashes on their own systems. Building shared tools is only valuable if less-capable partners can use them. It is no surprise that Meta has subsequently open-sourced a hashing protocol and is releasing an open-source system for maintaining internal hash databases on a platform's internal systems.[4] Innovative tools are meaningless unless they connect practically to the tools and systems needed to deploy them.

Conclusions

Adversarial innovation is dirty business. When the stakes are high, innovation is dangerous. The positive impact is rarely immediately clear, and it will produce new modes of error. Inaction is often less risky for individuals in a bureaucracy but poses more dangers to a long-term mission against an adaptive adversary. There is no greater lesson from Trust & Safety than that cycles of adversarial adaptation occur faster today than ever before.

Based on my experience in Trust & Safety, Commands should consider a variety of practical steps to enhance innovation:

Expect obsolescence. Innovation in an adversarial setting is never done. Expect that every process, technology, and framework will become outdated. Iteration and innovation happen incredibly quickly online because the cost of failure for attackers is low. But this dynamic exists elsewhere, and it is accelerating in many areas of military conflict. The cost of experimenting with new drone techniques is lower than with manned aircraft. Electronic warfare systems can be deployed, deprecated, and updated quickly by a determined adversary.

Hire unusual talent. Talent is destiny in technology. Find the introverts, the folks with blue hair, the ones who can rebuild an engine from scrap, and the people who are skeptical of working with the government. Show them that the mission matters and set them loose. Many of these people will not live in Tampa. Build Centers of Excellence in New York, Los Angeles, and San Francisco. That's hard for the government, but that is not an excuse. It is also hard in the private sector. OpenAI originally wanted all hires in its San Francisco office. But all the talent they needed was not where they wanted it, so they had to open offices elsewhere. If innovation is a top priority, the government must position itself to hire innovators where they live.

Build innovation around real problems. Generic innovation centers will not work to develop applied solutions. Applied innovation requires proximity to and responsibility for real, meaningful missions. Some missions are not well-suited to risky innovation, but you cannot de-risk entirely and expect new ideas. To that end, give your innovators real, practical problems. Assign responsibility for a critical mission to that innovation center - or simply demand innovation from a unit assigned a particular problem. You cannot innovate in a vacuum; you must feel pain and failure and risk to do it right.

"Talent is destiny in technology. Find the introverts, the folks with blue hair, the ones who can rebuild an engine from scrap, and the people who are skeptical of working with the government. Show them that the mission matters and set them loose."

Cross-functional organizations innovate better. Give an innovative team what it needs to try new ideas by embedding appropriate cross-functional resources within it. Do not make them beg a bureaucracy for resources and expect them to innovate quickly. Unleash this cross-functional team from dependencies on service-provider organizations, including by decoupling personnel seconded to that team from traditional rating processes.

Align strategic and operational leaders. Operational flexibility and dynamism are critical to success. Strategic leaders will rarely have the right answers; even dynamic, expert operational leaders must primarily empower bottom-up ideas within their teams rather than drive it top-down. How do you do this? Hire non-traditional operational leaders, empower them by emphasizing the importance of their mission and resourcing their efforts, and offer grace if (when) they fail productively. If your operational leaders learn and adapt quickly from failure, embrace that effort. Do not disincentivize experimentation by punishing failure and risk-taking. Expect that operational leaders with a healthy disregard for standard operating procedure will innovate more effectively.

Prioritize mission, not process. Crisis is useful because it creates urgency around the mission. At its most basic, innovation is what occurs when a mission is given primacy over an established process. This is why innovation is fundamentally disruptive to an organization: If it is not painful, it is not systematic. It is possible to empower innovation in sub-units of an organization, but to do so, strategic leaders must emphasize the imperative of their mission and offer the leverage to upend the process in order to achieve it. Expect this to be unpopular in other parts of the organization.

Better tools enable new process. The limitations of existing tools regularly shape the operational processes of organizations. They destroy creativity. Fight this every day. Imagine an optimal process to advance your mission—and envision the tools that would facilitate that reality.

Innovation exists throughout the stack. Innovation is not always sexy. The most important innovations do not necessarily occur at the point of the spear where action is taken against an adversary. Understanding that technological innovation is inextricably tied to process change helps illustrate the links between upstream changes and mission outcomes. Both strategic and operational leaders must understand the entire chain of information gathering, decision-making, and execution that leads to positive outcomes in order to prioritize the most impactful innovations.

Use tools that facilitate innovation. Tools (and contracts) that lock you into specific operational processes impede innovation.

Emphasize core tooling that can be reconfigured quickly for various roles and missions, and that can operate as a platform for time-bound or experimental efforts. As a practical matter, this means tooling that can be configured easily by non-technical staff and makes data easily accessible for use with new tools and processes. Tools that lock-in data impede innovation and undermine your mission. The companies that sell them are prioritizing their revenue rather than your mission. Do not use them.

Do not assume human-driven processes are more accurate. Automated systems have shortcomings, but modern AI regularly beats human decisionmakers at many scaled tasks. Expect automation to make unpredictable errors and consider when such mistakes are acceptable to your mission. But do not assume that human beings will be better in the aggregate. Measure both and compare.

Collaborative innovation often just means sharing the basics. Collaborative work in coalitions is incredibly difficult—and the political hurdles to cooperation are often more important than the technical elements. A key lesson is that collaboration is not just about creating a shared resource; it is also about ensuring that every collaborator is able to effectively use that shared resource. This seems obvious, but it is an easy mistake for highly resourced organizations working with less capable entities.

It is an age-old question: Does art imitate life, or does life imitate art? An updated version might ask: Does digital conflict imitate real-world conflict or does real-world conflict imitate digital conflict? The answer, of course, is that these processes are bidirectional, symbiotic, and deeply intertwined. But if the digital conflict managed by Trust & Safety teams has lower stakes, on average, than real-world conflict, it also faces a faster pace of innovation because the costs of iteration are lower. The most successful Trust & Safety teams embrace this challenge. They cannot match their adversaries' pace, but they can get faster, shape the digital terrain, and use myriad other advantages to achieve their mission.

Innovation is what happens when the mission really, truly comes first. Not an existing process. Not long-standing culture. Not bureaucracy. That is why building innovation around a real, critical mission is central to success. Technological innovation should drive process and decision-making changes. That likely means pain for someone in the organization. Managing and overcoming the prevarication that pain will engender demands leadership—humble, audacious leadership. CTC

Citations

1 Josh Boyd, "In Community We Trust: Online Security Communication at eBay," *Journal of Computer-Mediated Communication* 7:3 (2002).

2 "Facebook Community Standards Enforcement Report – Q2 2024 Report," Meta, August 2024.

3 See "Dangerous Organizations: Terrorism and Organized Hate" in "Community Standards Enforcement Report – Q2 2024 Report."

4 See https://github.com/facebook/ThreatExchange/blob/main/README.md