Inauthentic Behavior in Online & Digital Systems

Final Conference Report

THE UNIVERSITY OF CHICAGO

HarrisCPI
CYBER POLICY INITIATIVE
THE UNIVERSITY OF CHICAGO

Version 1.2 (11.01.20)

Table of Contents

Executive Summary

The University of Chicago's Harris Cyber Policy Initiative hosted a two-day conference on September 25th and October 9th, 2020 to examine Inauthentic Behavior in Online and Digital Systems and to examine the potential utility of launching a National Science Foundation Convergence Accelerator on the topic.

Inauthentic Behavior (IB) is naturally situated at the intersection of several disciplinary boundaries: computer engineering, education, law and regulation, social science, and behavioral psychology. This research agenda is well suited for an NSF Convergence Accelerator, which is structured to breed cross-disciplinary collaboration. A Convergence Accelerator can transcend silos and help incentivize a robust research agenda. By using a cohort structure, NSF can instigate research along a range of avenues with complementary outputs.

This report details the major sub-themes that form the basis for interdisciplinary research into IB. Next, the report highlights parallel research agendas in inauthentic behavior in a range of fields, including medicine and health, banking and finance, biology, marketing, education, sports, and the intelligence community. It also delves into some of the central lessons that can be drawn from other disciplinary approaches to IB. The report then profiles important potential partners and underscores the importance of horizontal and vertical integration that crosses disciplines and promotes data-sharing into IB.

Finally, the report profiles three main types of deliverables that would likely emerge from a Convergence Accelerator: educational models, tools for users, and tools for platforms. These solution sets are viable pathways for reaching tangible outcomes in the following two to three years, while simultaneously structuring longer-term research agendas in the field.

Inauthentic Behavior: A Proposed Convergence Accelerator

Inauthentic Behavior (IB) is a relatively new term employed by tech platforms, internet communities, and a growing number of researchers to describe fake, manipulative engagement by individuals and groups in online environments. Facebook coined the term in 2016 to structure its evaluation and enforcement methods for false and misleading content on its pages, particularly in the wake of the 2016 U.S. Presidential Election. The definition focuses on the activity itself, rather than on the specific content circulated or the actor behind it.

The platform is most concerned with "coordinated inauthentic behavior," or influence operations that seek to "manipulate public debate for a strategic goal where fake accounts are central to the operation."[1] However, inauthentic behavior can and does include financially motivated activities such as spam and artificial amplification.[2]

Research focused on inauthentic behavior naturally sits at the intersection of several fields, including computer engineering, education, ethics, law, and psychology. Any potential research agenda is plagued with several nettlesome methodological questions, including how to define and scope the intentions of inauthentic actors, how to quantify the impact of inauthentic behavior, and how to align incentives across industry and government to reduce harmful activities. IB research is well suited for a convergence accelerator approach, specifically because the field draws from a variety of disciplines and must be structured to account for gaps across technical, legal, and regulatory environments in order to lead to successful solution sets.

Sub-Themes: Education

Education efforts have arisen to try and tackle the rising problem of misinformation and disinformation online. Digital literacy, or "the ability to use information and communication technologies to find, evaluate, create, and communicate information requiring both cognitive and technical skills," is a particularly important tool to combat IB.[3] Digital literacy campaigns seek to equip individuals with the skills necessary to evaluate and avoid IB and other pernicious behaviors online.

Digital citizenship campaigns also help emphasize responsible use of information technology and help underscore the importance of maintaining a healthy online community that mirrors the national community. Digital citizenship "refers to the responsible use of technology by anyone who uses computers, the Internet, and digital devices to engage with society on any level." A digital citizen uses the internet "regularly and effectively" and can use online platforms to "connect with [others],

---

[1] https://about.fb.com/wp-content/uploads/2020/05/April-2020-CIB-Report.pdf

[2] https://about.fb.com/wp-content/uploads/2020/05/April-2020-CIB-Report.pdf

[3] https://www.renaissance.com/2019/02/08/blog-digital-literacy-why-does-it-matter/

empathize with each other, and create lasting relationships through digital tools."[4] These campaigns often incorporate digital literacy into a broader understanding of user data, digital divides, and practicing empathy in online engagement.

These campaigns often build off of research into education systems which highlight disparate digital access and ability. For example, Raj Chetty has worked on education and income/wealth inequality, highlighting the different socioeconomic factors that can influence a person's digital comfort. Understanding the existing gaps in digital literacy can help structure approaches to combat IB by focusing resources into the most vulnerable pockets of populations.

Sub-Themes: Psychology

Psychology offers a salient entry point for understanding the impact of IB on authentic actors who may engage with these coordinated efforts online. How are average digital citizens interacting with and perpetuating inauthentic behavior that originates in coordinated campaigns? In other words, do regular citizens become participatory members of IB communities? How do IB campaigns seek to solicit involvement from others online?

Individuals are often attracted to information online which confirms their biases and conforms with their existing worldview. Coordinated IB thus often seeks to accelerate certain understandings and disinformation, perpetuating false or misleading claims in communities which are already predisposed to accept these frames. Understanding how these biases form and how IB operates to manipulate existing gaps in understanding and proclivities toward conspiratorial thinking can provide useful baselines for thinking about the operational realities of coordinated IB.

Behavioral psychology also can offer useful starting points for understanding how to mitigate the effects of IB. Theories of nudging illuminate how small interventions can lead to tangible changes in human behavior. Assumptions of system design also often lead to key gaps that allow for the perpetuation of false or misleading information. Can such tactics be adapted to combat IB? What are the mechanisms that might mediate between IB and changes in belief or action downstream? There is also an interesting intersection point here with the growing field of the psychology of security, which could help elevate certain potential solution sets over others when adapting programs to deal with IB.

Sub-Themes: Political and Social Science

Political science has developed a number of methodologies to gather data on the effect interventions have on human behaviors. Surveys and survey experiments are useful tools for tracking individuals' attitudes and actions after a specific intervention. Such methods could provide useful insights into the impact of education and digital literacy campaigns on future engagement with IB. These survey experiments could also help elucidate how differences in individuals (e.g., age, skill set, cultural setting, education level, gender) correlate with susceptibility to IB or subsequent activities related to IB.

---

[4] https://www.aeseducation.com/blog/what-is-digital-citizenship

Social science also has a long history of dealing with the difficulties of measuring impacts. For example, a robust literature on political communication and voting patterns offer insights into how to approach measurement in patterns of behavior that are heavily mediated by other external factors and events. What is the *actual impact* of inauthentic behavior? Are there salient connections between online behavior and offline political activities? Social science research agendas can help structure research agendas on these important issues.

Sub-Themes: Computer Science

IB is not easily defined; many platforms have different standards for what qualifies as coordinated inauthentic behavior worthy of being tracked or disrupted. Without a consistent strategy for combating IB, platforms may leave too much discretion up to individual coding teams, leading to biases in enforcement. These issues once again reinforce the importance of developing a clearer understanding of what constitutes IB across online communities.

Many tech platforms have developed algorithms to identify and remove inauthentic behavior. Such solutions have had limited success, particularly against coordinated IB originating from sophisticated sources. Although platforms can often flag spam or fake advertisements, financially motivated IB tends to be less carefully constructed and easier to spot through traditional algorithmic models. Coordinated IB, particularly campaigns backed by nation states or well-financed non-state actors, is often incredibly sophisticated and quickly adapts to algorithmic detection efforts. Companies often rely too heavily on artificial intelligence and machine learning as catch-all solutions; however, as soon as malicious actors understand how these systems work, they shift their tactics to avoid detection. An over-reliance on algorithmic solution sets often leads to escalating gamification of tech platforms, where coordinated IB actors continue to change their targets and strategies to throw systems off their scent.

Other frameworks are also under development, including technosocial engineering, which focuses on how persuasion and influence operates online, and how technological tools affect individual decision-making and perceptions.

Sub-Themes: Mathematical and Statistical Modeling

Modeling efforts can also unpack the complex interconnections across vast digital communities, many of which span platforms, countries, and even languages. Visualization projects can help researchers evaluate and reexamine existing assumptions about how information travels in these communities, how ties are formed across communities, and how activity patterns evolve over time.

Kristina Lerman, Project Leader at the Information Sciences Institute at the University of Southern California, is using mathematical frameworks to model the collective behavior of social web users and to pinpoint developing trends. Lerman's work attempts to understand the feedback systems between individual and collective decisions.[5]

[5] See Kristina Lerman's work: https://www.isi.edu/integration/people/lerman/overview.html.

Sub-Themes: Law and Regulation

IB also touches a range of legal and regulatory challenges faced by tech platforms, national governments, and international institutions. How can regulation be used to align financial incentives on platforms with the reduction of coordinated IB? Many IB campaigns are able to operate because they have taken advantage of existing incentive structures which privilege quantified engagement such as likes and shares, even if that engagement is false, overqualified engagement such as interactions with real, verified people.

Tech platforms also operate across countries, leading to information sharing and enforcement challenges as states adopt new procedures for tackling the spread of harmful IB campaigns within their own borders. Legal challenges will continue to be at the forefront of any attempts to curtail damaging IB.

Convergence Potential

As illustrated in the previous section, any robust research agenda around coordinated inauthentic behavior in online communities must draw from a range of disciplinary and methodological approaches. IB also exists in many other systems; by examining parallels across other industries and sectors of society, we can learn important lessons that can be applied to current challenges in online IB.

Parallels: Medicine & Health

Medicine has long dealt with inauthentic behaviors in virology, immunology, and elsewhere. Parasites and viruses operate by pretending to be authentic parts of the human body, slipping in undetected and then wreaking havoc on internal systems. Cancerous tumors often grow slowly and mask themselves as regular cell growth, thus avoiding detection for longer periods of time. Health care systems have had to invest in sophisticated screening processes and testing procedures to root out dangerous intruders or abnormal processes within the human body.

Similarly, immunology seeks to equip the body with detection and prevention systems against harmful pathogens and infections, which often enter the body through inauthentic behaviors. Vaccinations are developed specifically to teach the body how to protect itself against a threat before it faces it; by introducing a small dose of a potentially harmful pathogen, vaccines help the body practice and strengthen its defenses in order to avoid falling ill in the future if exposed in the wild.[6] The very nature of a vaccine is predicated on IB, in a way — the body is exposed to a seemingly harmful substance that has in fact been engineered to be harmless, a teaching tool used to build up defenses against a future bacterial, viral, or parasitic invasion.

Parallels: Banking & Finance

Banking systems have to ensure authentic actors when operating systems of credit. Most purchases are made through credit cards, debit cards, or checks — such purchases essentially constitute a promise of payment behind the scenes, rather than the actual exchange of cash money at the time of sale. Banks thus have to quickly root out and remove inauthentic actors to preserve the trust of the existing system. Fraud detection systems are robust in modern banking and operate in conjunction with insurance schemes to shut down stolen identifications quickly. Tax schemes have also had to root out systemic cheating and falsification of certain benefits, such as unemployment or child credits.

Financial systems also face other systemic challenges, such as money laundering, which operate through filtering illicit money gains into credible institutions. Many of these systems are sophisticated, and banks have had to become quite adept at sorting out ill-gotten gains from legal profits. Here, incentives have had to be realigned through regulation: banks generally gain from taking in profits, but they face legal risks if they allow illegal gains to enter their systems.

---

[6] See https://www.immunology.org/public-information/what-is-immunology.

Parallels: Biology

Natural systems also deal with inauthentic behavior, often in the form of mimicry. A subset of camouflage, mimicry occurs when an animal (the mimic) pretends to be another organism (the model) in order to deceive a third animal (the target).[7] Mimicry is predicated on misidentification: although mimicry can be used to lure prey or to avoid predation, mimicry operates through active deception and obfuscation.

Like IB in the online realm, some forms of mimicry are harmless, or even mutually beneficial to both the mimic and the target. Müllerian mimicry is a collaborative form of behavior whereby two well-defended (e.g., poisonous or otherwise harmful) prey share similar warning signs. Predators need not learn additional signals of which to be wary, and prey benefit from being doubly protected.[8]

Other forms of mimicry, however, are quite costly for the targeted animal. One such type is brood parasite behavior, whereby one organism tricks another into raising its young. Although practiced by a range of animals, the behavior is most often associated with birds. Several species of bird have evolved eggs which look nearly identical to a potential host bird, thereby allowing them to offload their own eggs into host nests. The hosts then take care of the intruder eggs as if they were their own young, building nests and caring for unhatched eggs, a time-consuming and resource-intensive endeavor. Brood parasites also often spread their eggs across different host nests, thereby increasing the likelihood that some of their eggs survive. Brood parasites have also led to an arms race of sorts, as hosts develop new ways of detection to root out intruder eggs and would-be parasites develop more closely mimicked eggs.[9]

Parallels: Marketing

The marketing world has also dealt with issues of financially motivated deception and inauthenticity. Marketers have often skirted the line between exaggeration and flat-out falsehood. Legal regulations have evolved which seek to reign in how deceptive marketers can be — in other words, defining the line between puffery and lie. Parsing behavior is difficult, and courts often have to deal with issues of overzealous marketing. However, the system has established a framework of standards which are generally held up by traditional advertisers.

Parallels: Education

Education systems also have to deal with plagiarism and cheating. Many institutions have turned to automated platforms which check whether an essay or examination has lifted material from online materials. Plagiarism.org offers educators a host of resources designed to detect plagiarism, including

---

[7] Animal Behavior Dictionary.

[8] https://www.oxfordbibliographies.com/view/document/obo-9780199941728/obo-9780199941728-0062.xml

[9] https://www.cell.com/current-biology/fulltext/S0960-9822(13)01031-2.

Document Source Analysis, a tool found on Turnitin.com. This technology "works by assigning a unique identifier (called a 'digital fingerprint') to every text document". The tool helps detect not only direct copying, but also more sophisticated forms of plagiarism.[10]

Yet while technology has helped to catch plagiarists, it is an imperfect system. Many students plagiarize by accident, misunderstanding how to incorporate sourced material into an original argument. Resources such as Plagiarism.org also seeks to educate students about what constitutes plagiarism and how to properly cite materials when conducting research and analysis. Other students may figure out how to game the algorithmic systems of tools such as Turnitin.com, thereby avoiding detection at least in the short term. Platforms must continually evolve to stay ahead of students.

Parallels: Sports

Professional sports leagues have also had to deal with inauthentic behavior in the form of cheating. Referee systems deal with two primary groups of problems: cheating to benefit one's own performance or team and throwing a competition intentionally in order to benefit through secondary markets such as gambling rings.

Cycling competitions such as the Tour de France have struggled with doping scandals, including the use of stimulants, steroids, and hormones.[11] As with other cycles of inauthentic behavior, doping techniques and detection systems have coevolved, as international and state systems attempt to root out cheaters from competitions and competitors become more and more sophisticated in their doping techniques. For example, competitors have adopted systems of autologous blood doping, whereby they get transfusions of their own stored blood.[12] This type of doping is particularly difficult to detect, and official methods have been slow to develop.[13]

Other forms of inauthentic behavior have plagued the game of baseball. The Houston Astros were recently embroiled in a scandal involving sign stealing. Sign stealing is not automatically against the rules of professional baseball: in fact, teams are often praised for their ability to detect signaling systems by opponents. However, the MLB has banned the use of technology to steal signs. Although electronic communication and video systems have entered baseball stadiums, neither system is supposed to be used for the purpose of stealing signs. After a thorough investigation by the MLB, the Astros were found to be using replay monitors to learn the signaling systems of other teams.

Throwing competitions is also a familiar form of inauthentic behavior in sports. The 1919 Chicago White Sox scandal is frequently cited as one of the most significant examples of such behavior. The White Sox were accused of throwing the World Series against the Cincinnati Reds as part of a massive gambling

---

[10] https://cs.stanford.edu/people/eroberts/courses/cs181/projects/2000-01/honor-code/tech.htm

[11] https://www.cyclingweekly.com/news/latest-news/encyclopedia-of-doping-74006.

[12] https://www.wada-ama.org/en/questions-answers/blood-doping

[13] https://pubmed.ncbi.nlm.nih.gov/22407819/

ring. The responsible players, including “Shoeless” Joe Jackson, were banned from the game for life and became ineligible for the Hall of Fame.

Parallels: Intelligence Community

Intelligence agencies have also long dealt with issues of inauthentic behavior. Human assets often have to try and determine if the information they are receiving is reliable, or if instead they are being intentionally misled by sources. Technological aids have limited success: polygraph testing remains controversial, with some institutions suggesting that the machines are not much better than chance.[14] Humans are also notoriously bad at detecting liars. In fact, some studies have suggested that people who think they possess an intuitive sense of such matters are actually worse at lie detection than those who question such gut reactions.[15]

Intelligence agencies have worked to build systems to defend against such kinds of manipulation, including burn notices that work to remove any and all intelligence that originated from a known fabricator. Other methods include robust verification systems, whereby information must be verified through multiple sources before being trusted.

Lessons Learned: Verification

By examining parallels in these diverse fields, we can gain insights into how other systems have solved for inauthentic behavior.

Many systems build in multiple checks for determining whether an actor is a reliable source of information. For example, anti-doping schemes often depend on continuous testing regimes so that organizations can develop a profile of an individual against which to compare for possible artificial enhancement during competitions. Intelligence agencies flag information with confidence levels based on the number and reliability of sources and methods. Banks, too, rely on multiple forms of identification in order to open up accounts and make major transactions; behavioral markers are also used to determine whether a person is who they say they are, or whether someone’s credentials have fallen into the wrong hands. Vouching systems also work off of similar assumptions: identities are corroborated through validation systems that process documentary evidence and rely on several authorities.

Verification processes have been adopted in many tech platforms in order to combat inauthentic behavior, with varying degrees of success. Whitelisting approaches to identity can help reduce false engagement. However, verification systems can often lead to gamification, whereby malicious or fraudulent actors work to continually trick system innovations to skirt detection. These systems can also create high barriers to entry for online platforms, reducing the number of eligible users and stifling engagement.

---

[14] https://www.apmreports.org/story/2016/09/20/inconclusive-lie-detector-tests

[15] https://www.newscientist.com/article/dn2054-intuitive-people-worse-at-detecting-lies/

Lessons Learned: Incentives

Realigning incentives can also help to reward authentic behavior and punish inauthentic behavior. Authentic behavior is rewarded, and inauthentic behavior is punished. Market feedback loops have helped to reign in falsehoods in advertising: if customers realize that they are being scammed, then they will abandon that product and opt for something more credible instead. Financial systems have also built-in costs and benefits that reward authentic behavior and disincentivize inauthentic activities, including credit card fraud and money laundering. Individual consumers are generally not liable for credit card fraud if caught and reported quickly, meaning that liability falls on merchants and card issuers. Indeed, merchants are often held liable for fraud in cases of card not present transactions, which constitute essentially all online and telephone ordering systems.[16] Thus, companies have strong incentives to reduce fraudulent transactions since they have to absorb many of the costs associated with these inauthentic behaviors.

Anti-doping and anti-cheating systems in sports often involve severe punishments for violators, including lifetime bans from competitions. These can help to push actors to rethink the benefits of inauthentic behavior. The costs associated with biological inauthentic behavior can also help to reduce the practice. For brood parasites, the potential costs of being discovered are quite high, generally involving destruction or abandonment.

Tech platforms often implicitly reward inauthentic behavior. The practice of purchasing likes or shares on Instagram and elsewhere continues to proliferate, particularly as trending topics continue to be based on such metrics. Many users cannot easily decipher real interest from fake propagation, and platforms do not always catch such false behaviors quickly enough to stop their spread. Realigning incentives to reward authentic behavior is a necessary but difficult step to combatting coordinated IB.

[16] https://www.signifyd.com/resources/fraud-101/why-liable/

Partnerships

IB is an issue that benefits from a multidisciplinary approach, and from engagement across traditionally segmented communities.

Partners: Technology Companies

Tech companies are an obvious and critical partner in understanding and combating IB online. Major platforms such as Facebook and Twitter have had to adapt to coordinated IB campaigns by both state and non-state actors. These platforms continue to institute new policies aimed at reducing false or misleading activities. These companies have a clear incentive to reduce spam or fake advertising, which might lead users to abandon their sites. However, they also have to be careful in how they mitigate other kinds of IB. Many IB campaigns touch on politically sensitive issues like free speech; the removal of certain kinds of messages is often construed as politically biased, particularly if those messages highlight or stoke partisan issues.

Tech companies are often hesitant to share data with government actors for fear of regulation or litigation. Information-sharing efforts across tech platforms is often easier to achieve than across platforms and other industries.

Partners: Media Companies

Media companies are often plagued with issues of IB. News outlets are often flooded with fake comments that seek to discredit published articles, or even offer threats against specific reporters. These companies also have different incentive structures in place than traditional media platforms. Rather than valuing pure engagement, media companies want users who engage with their information because they find it trustworthy. Media companies could help experiment with new ranking systems that privilege certain users over others based on how “authentically” those users engage with their content. In other words, users could be scored based on their own behaviors and ability to parse fact from falsehood; engagement by higher ranked users would be weighed more heavily than engagement by more gullible users.

Media companies could be a useful partner in developing alternative methods of defining engagement online based not simply on profitable clicks and shares, but on trusted interactions with digitally literate users. Experiments in these areas could help bolster efforts across other, broader tech platforms like Facebook and Instagram.

Partners: Non-Governmental Organizations

There are a range of NGOs whose work is relevant for understanding and combatting IB. Censorship and media literacy NGOs both in the US and in other countries could help provide frameworks for balancing freedom of expression with the removal of IB. Freedom House, Media Matters, and other organizations could be a useful starting point.

Think tanks and institutions that focus on policy-relevant research could also provide an important bridge between industry and government. Many think tanks in the US and around the world have expanded their focus to include digital governance, influence operations, and disinformation campaigns. These organizations often have robust data sets that could help inform qualitative and quantitative research into IB.

Partners: Government Institutions

Governments are also important partners in this space. Governments shape regulation and law which defines how tech platforms can operate, and will need to be involved in conversations around how best to effectively leverage such tools to combat IB. Over and under-regulation are problematic: while overregulation could stifle innovation and restrict free speech, under-regulation has led to the current environment of IB proliferation. Tech platforms are generally left to police themselves, a rarity in any industry. Legislative and regulatory organizations like the Federal Trade Commission, the Department of Justice, and the Securities and Exchange Commission can help incentivize platforms to operationalize tools that stymie the spread of IB.

Partners: Academic Community

Academic institutions often have robust research programs on data regulations and protections, internet usage and trends, and cognitive and behavioral sciences, all areas which help inform IB research. A number[17] of universities have dedicated research hubs which focus on policy-relevant outputs. Stanford's Freeman Spogli Institute for International Studies (FSI) has launched a Global Digital Policy Incubator, a platform dedicated to advancing policy solution sets. Similarly, Harvard's Shorenstein Center on Media, Politics and Public Policy has a project on Technology and Social Change which focuses on understanding how manipulation of the media erodes democratic societies.[18] Academic centers can help curate data and define measurement tools for research and development into IB.

Considerations: Verticals vs. Horizontals

Integration is necessary both within and among communities that deal with IB challenges. Many research and development efforts focus on bringing partners together from disparate industries. Tech platforms have been traditionally reticent to engage with government actors on such issues out of fears of added regulation; however, neither sector can effectively combat IB alone. Similarly, state governments often avoid interaction with international organizations because of diverging interests or concerns: some states might have narrower understandings of free speech, for example, and thus may desire stronger enforcement mechanisms than other states. However, such differences should not hinder the development of some common frameworks and baseline understandings of what constitutes troubling IB.

---

[17] https://cltc.berkeley.edu; https://fsi.stanford.edu/front-global-digital-policy-incubator

[18] https://shorensteincenter.org/programs/

While horizontal collaboration is important, vertical integration is also essential. Many industries involved in IB enforcement are deeply hierarchical; information and practices become siloed and are not communicated and shared within organizations or communities. Law enforcement, for example, has strict jurisdictional requirements that can often slow down information sharing and coordination. Similarly, in the tech community, many IB detection procedures are buried in departments and are not evaluated strategically as a company or community-wide issue.

Considerations: Data Sharing

Data access continues to be a problem for IB researchers. Tech platforms have well-established data sharing programs, particularly to alert one another of new threats they unearth. However, raw data on user behavior is highly controlled by platforms since such information is a primary source of revenue generation. Many tech platforms also carefully curate and interpret data before publicizing it, making it difficult for independent researchers to discern actual trends in behavior and enforcement. Data sharing standards need to be established in order to provide researchers with a more complete picture of what is going on in these communities, without overt framing from industry.

Partnerships can help ease these data sharing issues by incentivizing pooled resources and removing financial interests from the equation. The federally funded research centers (FFRDCs) model useful: FFRDCs frequently use classified or sensitive information that could easily be abused. Researchers at these organizations undergo security clearance processes, and their findings are carefully assessed before they are disseminated in any kind of public forum. Similar data sharing communities could be established, allowing industry and researchers to share highly sensitive information with one another in the context of combatting a shared threat: inauthentic behavior.

Considerations: Cross-Functional Focus

Most importantly, partnerships should be structured to encourage collaboration across traditionally siloed domains. Professional incentive structures vary significantly from industry to government to the academic community; often, work that crosses boundaries is underfunded at the organizational level and undervalued in individual career evaluations. The NSF Convergence Accelerator can provide a necessary framework for pushing industry experts and academic researchers out of their comfort zones and into a fruitful multi-disciplinary environment. Cohort funding structures could also help push forward a range of different projects with interesting and complementary goals.

Deliverables

A potential NSF convergence accelerator track focused on inauthentic behavior should look to deliverables across education programs, tools for users, and tools for platforms in order to carry out the themes described above. These potential deliverables are primed with premier partnership opportunities that were detailed in the prior section.

Education Programs

Inauthentic behavior does not consist of bright lines, but rather many shades of gray. Educational programming is needed to highlight the forms IB can take, the potential intentions behind coordinated IB campaigns, and the ways users can protect themselves from spreading falsehoods online.

Digital literacy programs have been developed in many countries, including Latvia and Estonia, which seek to equip individuals with tools for evaluating the veracity of online information. Throughout the EU, Safer Internet Centres have emerged which provide tools for protecting children and young adults online.[19] Safer Internet Centres also launch awareness campaigns that highlight some of the risks of malicious or false information online, and help build societal capacity for combatting IB.

Programs such as CSforAll highlight authenticity as part of its curriculum. CSforAll seeks to provide students with coding and computer skills as early as elementary school; continuous exposure breeds familiarity and comfort.[20] Other educational programs have built interactive games which allow individuals to wade through information and practice evaluating its authenticity. Games can also help illustrate just how easy it is for a person to become an unwitting asset of IB.

The NSF convergence accelerator should invest in education programs that incorporate IB awareness at a young age. The accelerator should also support data-driven educational programming that targets vulnerable communities with easily digestible information about the risks of IB. Gaming and other interactive efforts could be particularly useful in this vein. Research and development into education tools likely touch on at least three NSF research areas: Research on Learning in Formal and Informal Settings (DRL) within the Education and Human Resources Directorate (EHR); Computing and Communication Foundations (CCF) within the Computer and Information Sciences and Engineering Directorate (CISE), and Behavioral and Cognitive Sciences (BCS) within the Social, Behavioral and Economic Sciences Directorate (SBE).

Tools for Users

Countering IB also requires investing in tools for individual users to better understand and avoid false or misleading information online.

---

[19] For more information, see: https://ec.europa.eu/digital-single-market/en/safer-internet-centres

[20] https://www.csforall.org

User agreements offer an interesting possible solution set. Currently, most user agreements are densely packed with legal minutia; most people do not invest any time in trying to understand them, clicking "accept" before they have even scrolled through the text. However, user agreements could be structured to be more accessible to the lay user, highlighting behavioral norms and clearly stating the way data is stored and shared. These agreements could be used as a helpful starting point for setting norms around responsible, authentic behavior.

Similarly, user codes of conduct could be a helpful tool for individuals navigating vast online spaces. Users could be exposed early on to a set of specific actions to take in order to verify information and report falsehoods. These codes of conduct could be structured around a set of quick and easy steps for evaluating information.

"Fact-checking" or information verification tools could also be a helpful tool for users. A plug-in or extension could seamlessly integrate with the user's preferred web browsing application. These kinds of tools would obviously not be infallible but could help users quickly discern how reliable a source may be. The tools could also help to prompt users to adopt digital literacy skills, flagging potentially problematic information to steer users toward further research on the topic.

The NSF should invest in tools for users that enable and reward responsible individual behaviors. These tools should help prompt better user engagement on the internet by highlighting possible problem areas and equipping users with control over their own data and activity online. Research and development efforts into user tools combine at least three NSF research areas: Information and Intelligent Systems (IIS) within the Computer and Information Science and Engineering Directorate (CISE); Research on Learning in Formal and Informal Settings (DRL) in the Education and Human Resources Directorate (EHR); and Behavioral and Cognitive Sciences (BCS) within the Social, Behavioral and Economic Sciences Directorate (SBE).

Tools for Platforms

Technology platforms currently have fairly weak defenses against IB, particularly sophisticated, coordinated campaigns. Platforms have been wary of restricting information on their sites out of fear of added regulation or scrutiny. However, platforms are now recognizing the importance of protecting their user bases from rampant disinformation and inauthentic behavior, lest they lose trust and relevance.

Platforms have a variety of ways to slow down the spread of information by introducing "friction" into their engagement tools. While platforms rely on rapid and frequent interactions, they also have an interest in maintaining the authenticity of those interactions. Platforms can consider slowing down the posting process by adding in extra checks. Twitter recently reframed the "retweet" option so that users were first prompted to "quote tweet" and add their own content and had to click out of such an option before simply resharing an existing post without adding their own thoughts. The change is very new and has faced considerable black lash, but it may help to avoid rampant retweeting of false information by building in a behavioral check that encourages users to consider their own take on the information first. The change is slated to stay in place until after the American presidential election, at which point Twitter

will evaluate its impact. Twitter is also exploring adding a "read before you retweet" prompt, encouraging users to make sure they have evaluated a source before sharing.[21]

Platforms can also introduce community moderation features. Communal verification and exclusion can operate as a mechanism for trust, helping to build smaller, more reliable pockets within larger platforms. Such mechanisms raise additional problems, including reinforcing hierarchies and privileges through exclusionary tactics. These small communities can also become echo chambers for dangerous beliefs; Reddit, for example, has become host to conspiratorial communities such as Qanon believers and anti-vaccine proponents through its subreddit capabilities. Platforms should explore ways of creating community moderation that emphasizes trustworthy behaviors. Such communities can be formed around preexisting belief systems, such as political affiliation or local interest. However, reputational systems could be a useful addition to strengthen moderation and removal of IB.

Platforms also need to reevaluate incentive structures so that the quality of engagement is rewarded, rather than the quantity. Current models of engagement are too easily gamed by inauthentic actors; although some platforms have become more adept at detecting fake amplification, users are all too easily coopted into sharing false information on their own. Platforms should consider weighing engagement based on the authenticity of the behavior. Here, reputation systems could also be useful in helping to segment users based on their ability to identify and avoid IB.

Platforms can also evaluate and expand whitelisting efforts, whereby certain users are certified and authenticated as trusted. These systems are imperfect; the Twitter "blue checkmark," for instance, has been awarded to falsely credentialed users.[22] However, they can be helpful in providing an additional datapoint for evaluating the source of information.

Most importantly, platforms should recognize that no system will be perfect: IB is a problem that will likely be with us for a long time. However, by adopting a "defense in depth" mindset, platforms can begin to tackle the spread of misinformation and falsehoods. Research and development into tools for platforms draws from at least three areas of NSF research: Computer and Network Systems (CNS) and Information and Intelligent Systems (IIS) in the Computer and Information Science and Engineering Directorate (CISE); Research on Learning in Formal and Informal Settings (DRL) in the Education and Human Resources Directorate (EHR); and Behavioral and Cognitive Sciences (BCS) in the Social, Behavioral and Economic Sciences Directorate (SBE).

[21] https://techcrunch.com/2020/10/09/twitter-retweet-changes-quote-tweet-election-misinformation/

[22] https://nakedsecurity.sophos.com/2013/01/17/twitter-fake-verified-account/