Why Jenny Can't Figure Out Which Of These Messages Is A Covert Information Operation

Tristan Caulfield
University College London
t.caulfield@ucl.ac.uk

Jonathan M. Spring
Carnegie Mellon University
jspring@sei.cmu.edu

Angela Sasse
University College London
a.sasse@ucl.ac.uk

ABSTRACT

We view foreign interference in US and UK elections via social manipulation through the lens of usable security. Our goal is to provide advice on what interventions on the socio-technical election system are likely to work, and which are likely to fail. Strategies that the usable security literature indicates are likely to work are those that (1) avoid overloading the user's primary task; (2) help people understand negative consequences of their actions; and (3) support the long-term education of users with analytic reasoning skills and adequate background knowledge. Several of the responses to election interference proposed by governments and technology companies so far do not abide by these recommendations and are likely to be ineffective.

CCS CONCEPTS

• **Security and privacy → Usability in security and privacy**; *Social aspects of security and privacy*; • **Information systems →** *Social networks*.

KEYWORDS

disinformation, information operations, election interference, usable security

ACM Reference Format:
Tristan Caulfield, Jonathan M. Spring, and Angela Sasse. 2019. Why Jenny Can't Figure Out Which Of These Messages Is A Covert Information Operation. In *Proceedings of NSPW '19: New Security Paradigms Workshop (NSPW '19).* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The ever-increasing connectivity afforded by the internet and the growth of social networking platforms has created an environment where information can be disseminated very rapidly. Such connectivity can be useful, allowing individuals to share their thoughts, feelings, and experiences with friends and family around the world; it can help people stay abreast of news and politics; it can simplify organization and coordination of communities and movements. The internet has also be used to mislead people—from mere hoaxes and rumours, to fake news stories designed to be salacious and generate advertising revenue, to organized campaigns to promulgate falsehoods—with consequential physical and political effects. We will follow Gelfert's definition of *fake news*: "cases of deliberate presentation of false or misleading claims as news, where these are misleading by design" [26]. A broader term of art is *information operations*; defined by the US Joint Chiefs of Staff as "the integrated employment, during military operations, of [information-related capabilities] in concert with other lines of operation to influence, disrupt, corrupt, or usurp the decision making of adversaries and potential adversaries while protecting our own" [38]. Fake news is one of many methods by which an adversary might conduct information operations.

Our new paradigm is the intersection of two proposed changes to the way we think about foreign election interference. The first step is to expand the organizing model of "cybersecurity incident," which in the current cybersecurity paradigm covers security policy violations in information systems only (per [54]), to include security policy violations in socio-technical systems (namely, foreign election interference). That is, we think there is value in considering election interference as a cybersecurity problem. If we can think of election interference in terms of cybersecurity, then we can also apply ideas from usable security — and this is the second step. We will demonstrate usable security is a useful tool for suggesting both technical and non-technical approaches to mitigating the problem of foreign election interference.

Elections in 2016, both in the US and UK, have focused attention on the integrity not just of technological voting systems but also the social voting systems of which the technical systems are a part. As a result, societies are increasingly viewing their elections and other democratic institutions as part of their critical national infrastructure—systems which must be secured against attacks on their integrity. In this paper, we take the point of view that social manipulation, especially via network technology, is a *security incident* [54] within the socio-technical election system. More explicitly, our scope is threats to elections in the US and UK; specifically social manipulation of the electorate via technology. This point of view helps give structure to analysis of the problem of election interference by metaphor to information security. We will explore how this point of view opens up insights from the allied discipline of usable security that inform what may or may not be a viable and usable security architecture.

1.1 Scope

The scope of election interference is diverse. We limit our scope to foreign interference in US and UK elections; such interference is explicitly illegal in both jurisdictions. Even so, the actions an adversary might take occupy multiple modalities and time scales.

NSPW '19, Sep 23–26, 2019, San Carlos, Costa Rica

ACM ISBN 978-1-4503-7647-1...$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

There have been attempts at compromising the computers voters use to vote, and the computers local governments use to tabulate votes. There have been attempts to increase motivation to vote in targeted segments of the population via propaganda, promoting certain news stories, creating rallies, creating internet memes [64], recruiting unwitting local nationals, releasing slanderous material, sowing discord, and general fear-mongering. Adversaries have also attempted to reduce motivation in other segments of the population in these ways. One could view the security goal as assuring the integrity of the information within the social sphere. However, even if an organization or government could assure the truth or trustworthiness of information items, we want to avoid creating a dystopian Ministry of Truth. Orwell provides sufficient argument against that path [44]. Thus the security assurance we seek is that actors within the information exchange environment are incentivized to act in good faith and in compliance with relevant laws and regulations.

To consider this from a cybersecurity point of view, we need a statement of the security policy which is violated in this case. We do not mean that all types of election interference necessarily takes place in cyberspace, but rather that our definition of what an incident is comes from cybersecurity standards and the possible responses we consider are informed by those used in cybersecurity. The US Federal Election Commission (FEC) provides a summary of the relevant laws (52 U.S.C. §30121 and 11 CFR 110.20), which states that

> FEC "regulations prohibit foreign nationals from directing, dictating, controlling, or directly or indirectly participating in the decision-making process of any person (such as a corporation, labor organization, political committee, or political organization) with regard to any election-related activities" [62].

Similarly, the UK Parliament explicitly recognized the threat of social manipulation aiming to influence elections:

> In common with other countries, the UK is clearly vulnerable to covert digital influence campaigns and the Government should be conducting analysis to understand the extent of the targeting of voters, by foreign players, during past elections. We ask the Government whether current legislation to protect the electoral process from malign influence is sufficient. Legislation should be in line with the latest technological developments, and should be explicit on the illegal influencing of the democratic process by foreign players [15, p. 71].

Therefore, our perspective is that election interference is a security policy violation if the security policy is understood as the national law, and the information system is instead understood as the socio-technical system of the election writ large. Then information operations by hostile nation states or fake news items are security incidents if and only if they violate the policy. The national regulators decide what counts as incidents when they specify the security policy (that is, the law). Therefore, we take as security incidents just those events that violate the legal policy referenced above. Security requirements (confidentiality, integrity, availability, etc.) are defined by the security policy, as usual [54].

The FEC summary makes clear that it is the decision-making process of every legal person that needs to be secured; the security requirements are straightforward to adapt to this context. Transferring the legal terms of the FEC summary, the prohibition on "directing, dictating, controlling, or directly or indirectly participating in the decision-making process of any person" is primarily an assertion of the integrity of the decision-making process, but denying the availability of a decision-making process would certainly be a kind of control over it. Appropriate confidentiality is also implied by these legal terms; persons might be punished or their votes bought if foreign nationals can violate the confidentiality of decision-making processes.

We approach election interference from the technical perspective of *usable security*. We are not, in any way, claiming this is the only viewpoint or primary viewpoint. We are simply adding some evidence-based constraints on a viable solution which the usable security literature has to contribute. Election interference by social manipulation is an extraordinarily complex problem; no single viewpoint will solve it. However, as countries hasten to implement various measures to secure their elections, we wish to advocate that the lessons of usable security not be forgotten.

1.2 Related work

We are not the first to consider a cybersecurity angle on election interference. Farrell and Schneier [17] consider the impacts of information operations on the assurance models of democratic and authoritarian governments. The main contribution is to argue that the two governing systems react differently to contested information environments. Contested political knowledge may stabilize authoritarian regimes, though it destabilizes democratic regimes, because it affects each regime type differently. We aim to take a different approach and analyze what interventions may reduce such contested knowledge in democratic regimes.

There is a new but quickly growing literature on different aspects of social manipulation through technology. A 2018 review of the relationship between "social media, political polarization, and 'disinformation' " raised more questions than answers, taking its broad scope to include almost all slanted or incorrect information about politics [60]. Automated bots have been put to various social purposes, from political spam during crises to boost politician's apparent popularity [63]. As of 2017, there were documented well-organized "social media manipulation campaigns" in 48 countries, including domestic and foreign manipulation [10]. The Poynter Institute maintains a list of actions taken by various countries to counter information operations and fake news [24], which gives a wider overview of what is being done around the world. Researchers have also studied how consumers search for, acquire, and reason about news they read online via social media, finding a wide variety of strategies and success rates currently in use [21].

Our position within this quickly growing space is aided by our comparatively narrow scope, focusing on foreign (that is, illegal) election interference in the US and UK. Furthermore, we will offer a framework for forming predictions about what types of interventions are more likely to work, rather than documenting problems. We focus on the social manipulation aspect of election interference,

and so do not discuss technical aspects of voting or voting machine security. While voting machine security is crucial, the academic computer security community has been engaged and we have nothing new to add on voting machines themselves. Electronic voting machines in the US have had well documented and copious flaws in the 2000s [9, 20] that have continued well into this decade [43].

Section 2 sketches the pros and cons of five types of interventions on the socio-technical election system that might increase assurance in the desired security properties. We analyze the usability of these types of intervention in a very broad sense. Section 3 takes specific examples of intervention plans and evaluates whether they meet specific recommendations drawn from the usable security literature. Section 4 offers recommendations for promising directions and where future work might be most needed.

2 INTERVENTIONS

This section introduces fives broad types of existing interventions on the socio-technical system around elections. The five types of interventions have technical, policy, and human components. Capabilities and incentives, of both economic and psychological types, are a recurring theme within these interventions. Section 2.1 discusses interventions under the general grouping of international relations. Section 2.2 discusses social manipulation from the perspective of technical interventions to limit abusive content (e.g., spam). Section 2.3 discusses what to do after detecting an incident of social manipulation. Section 2.4 considers future interactions between campaign finance regulations and cybersecurity and cybercrime. Section 2.5 provides a historical perspective on public education versus propaganda and highlights some current educational efforts.

Our treatment of the material is brief and suggestive. A complete coverage would take a book. However, we have selected our types of intervention and examples to cover a broad range of levels. We have international policy, domestic public policy, organizational policy, and technical policy aspects. Usable security provides lessons at all these levels. This section gives us a toehold at each of these levels, before Section 3 provides a more thorough usable security view of some specific proposals.

2.1 International Relations

Historically, states have handled disputes between sovereigns via the various organs of power studied under international relations: war, trade policy, treaties, etc. As in many other areas, the internet and information technology has disrupted this historical pattern. On the one hand, non-state actors have increased effectiveness and reach. On the other hand, state actors have multiple options for false-flag operations and avoiding attribution by working through or appearing as non-state actors.

The traditional acceptance of espionage as part of state-to-state interactions would appear to classify automated or remote information operations, such as socio-technical election interference, as another way a state may be expected to excercise power or influence over a rival. In oversimplified terms, states generally expect they ought to get away with whatever espionage they can. The defending state essentially must deter espionage, for example by catching the spies or interfering with their equipment. Both of these options are much less effective given a globally shared internet. The spies never need leave home. Countries such as the Russian Federation [52, §61] and the People's Republic of China [66, §8(1)] will not extradite any of their citizens to any country, let alone government employees to the US, regardless of whether the US has named those persons in indictments for espionage. Thus, catching the spies is almost entirely ineffective.

To interfere with the equipment that foreign actors use to interfere with elections, the defending nations (the US and UK in our case) would need to conduct offensive cyber operations. We expect this to be ineffective because the adversaries are using the same network as the target citizens—the global internet. The adversary's equipment cannot be totally disrupted without turning off the whole internet—an unacceptable option in the US or UK. However, Russia is reportedly planning to test disconnecting from the wider internet as a defense against broader cyberattacks [14] and India shutdown its internet in response to lynchings triggered by rumours spread over the internet and mobile messaging [22]. Degrading adversary capabilities may be a short-term stop gap measure for particularly important events. For example, there are reports [28] that the US targeted Russian information operations systems during the 2018 midterm elections.

However, in the long run, the adversaries will learn how to make their infrastructure robust to such attacks. Similar adversary capabilities have developed and become widely available on time scales of 2–5 years [58]. Furthermore, much of the interference with the socio-technical election system are distributed over many months before the election, such as spreading propaganda, sowing discord, and creating memes. Given that much of this activity happens on shared social-media platforms, the infrastructure the adversaries use for these tactics cannot be interrupted in the same way.

2.2 Technology platforms and anti-abuse

Another natural option for incident mitigation is the platforms which host important aspects of the attack. The platforms usually have stated acceptable use policies. Abuse is any use of the platform that violates these terms. The platforms use a variety of anti-abuse measures to prevent abuse of their systems. The Messaging, Mobile, and Malware Anti-Abuse Working Group (M^3AAWG) is an industry association centered around sharing such measures.

If we view the election system as a socio-technical system, then attacks on the citizens are in scope. Given the FEC statement, interference in the decision-making of any person is against what we are taking as the security policy. Any news outlet that carries or propagates information operantions or propaganda is therefore part of the infrastructure of the attack. Nation-states have long sponsored media outlets that propagate views contrary to the narrative rival nation-states propagate. Therefore, an infosec framework for interpreting attacks on the voting system could roughly talk about this as insider attacks and external attacks. Explicit propaganda outlets are external, and in general behave according to their editorial policy. We should address the threat posed by such external propagandists with the tools of international relations, as they've been managed in the past. In this section, we focus on the category in which the majority of mass media and media platforms fall, that of some sort of co-opted insider. Insider here just means

the media company's stated goal of existence is not to propagandize on behalf of a rival nation-state. Especially in the case of media platforms, this insider/outsider distinction is blurry because citizens of any state may participate in the platform equally. However, from the perspective of US elections, media companies domiciled in the US or other allied nation-states could be convinced or compelled to change their policies in a way that outside media cannot. Thus our distinction of insider/outsider refers narrowly to the media platform itself, not its users.

Mass media has shaped public discourse since its inception. The technology platforms now hosting the public discourse are just that—platforms. Platforms exert less direct control over the content they host and propagate than traditional mass media. Media platforms and traditional mass media share the goal of shaping public opinion, insofar as the platforms are advertising companies. However, the basic observation [32, p. xi] makes about traditional mass media companies may apply to media platforms:

> "…[an attempt] to explain the performance of the U.S. media in terms of the basic institutional structure and relationships within which they operate. It is our view that, among their other function, the media serve, and propagandize on behalf of, the powerful societal interests that control and finance them. The representatives of these interests have important agendas and principles that they want to advance, and they are well positioned to shape and constrain media policy. This is normally not accomplished by crude intervention, but by the selection of right-thinking personnel and by the editors' and working journalists' internalization of priorities and definitions of newsworthiness that conform to the institution's policy"

We use this quote mostly to note that substantial changes to the policies of media platforms will likely not be easy, as established interests may resist changes to a system currently benefiting them. With this caveat in mind, we consider several possible responses to the attack on the election socio-technical system drawn from experience fighting other kinds of abuse on media platforms.

By treating election manipulation as abuse of the system's intended purpose, we are essentially following Brunton's definition of spam: "spamming is the project of leveraging information technology to exploit existing gatherings of attention" [11, p. xvi]. However, Brunton's definition of spam is in relation to some defined community that values the attention of its members and does not want it wasted or exploited by outside members. The online platforms are comprised of multiple and various communities, without such clear boundaries. The communities of interest are at least reasonably well defined in the case of vote manipulation. The citizens of a given country are the community, whether or not they can or plan to vote, and citizens of all other countries are not (per the law, i.e., security policy) to manipulate that community. Of course, the nature of online platforms, even more than mass media before them, makes this distinction difficult to maintain in practice.

Any technology platform today has automated policies in place to limit spam—that is, to limit the unwanted exploitation of the attention gathered on the platform. We will analyze the extent to which these existing solutions to limit abuse and spam might be suitable for addressing attacks on elections. And many platforms have started to classify propaganda as a kind of abuse, that is spam. For example, Twitter representatives have stated Twitter is

> "committed to understanding how bad-faith actors use our services. We will continue to proactively combat nefarious attempts[, such as propaganda and information operations,] to undermine the integrity of Twitter, while partnering with civil society, government, our industry peers, and researchers to improve our collective understanding of coordinated attempts to interfere in the public conversation" [25].

Pushing the task of removing propaganda to automated processes and professionals is good from the perspective of usability.

There are multiple constraints on this solution. One constraint is that, in many cases, the people propagating the information are legitimate users. Norms and laws preventing censorship are in tension with anti-abuse activities to remove or prevent dissemination of content. Secondly, anti-abuse work is done on a best-effort basis, and in many important aspects depends on volunteers [37]. The economic incentive for the platforms is to reduce abusive content to tolerable levels, not eliminate it. The same will be true of fake news and propaganda. Finally, these platforms are advertising companies. Their essential function is to manipulate public opinion in the ways those paying them desire. In line with [32], it would be naïve to trust advertising firms to tackle the problem of election interference and propaganda voluntarily.

2.3 Detection and mitigation

Section 2.2 assumes that propaganda is detectable. Surely there are technical challenges here. But we trust that fields like sentiment analysis can be brought to bear with existing abuse detection methods. Various social media companies have announced removals of various bots or accounts detected as propagandists, which would tend to support the claim detection is possible, if not perfect. The behaviour of troll or bot accounts changes over time, which can increase the difficulty of automated detection [65]. However, all these technical discussions may mask important decisions about what mitigation means if humans have already been exposed to manipulative content.

Human brains have a tendency to stick to the first opinion they form about a topic. This phenomenon is anchoring bias [35], one of many cognitive biases relevant to flawed or manipulable decision-making [13]. Incident response to a social event such as a news item therefore cannot follow all the norms for responses to computer security incidents. For example, removing a story because it is "fake news" is not the same as blacklisting a domain name or IP address. Unlike a piece of technological infrastructure, the human brain can not simply be reset to a sound state—thoughts and ideas persist, even when evidence to the contrary is presented.

Because of various cognitive biases, it may actually be harmful to mitigate a fake news story by tagging or removing the content. At best, tagging news items with warnings has a minimal positive impact on user perception of those items while conferring a false sense of security on false news items that are erroneously not tagged [46]. For some sub-populations, warning tags increased user belief. And it is certainly the case that technology platforms

will not correctly classify all news items. An unavoidable issue with the attempt to tag news as political or false is that "[n]obody, including Facebook, wishes this organisation to have such a level of control over the free press or even political campaigning" [6].

Some anti-abuse measures prevent content from ever reaching human eyes. But some inevitably slips through. In the context of domain names, reactive blacklisting alone cannot remove the incentive for the adversary to attack. Even if the adversary only gets a benefit once in a million tries, if they pay no cost but stolen time on compromised machines, they still profit [57]. The same dynamics likely apply to abuse with the goal of election interference. Since some attacks will continue to get through, and the attacks are likely to keep coming, communities and technology platforms should think about what a successful mitigation of the harm done by a propaganda campaign looks like. A basic step during computer security incident response is mitigation. What would be an adequate mitigation plan for attacks on the social sphere is unclear.

One option to change the dynamics is to punish the party perpetuating the abusive material or propaganda. In France, legislation passed in 2018 to reduce false and misleading information during election campaigns allows judges to order that articles online be immediately removed. The law also allows television channels controlled by a foreign state to be suspended if they 'deliberately disseminate false information likely to affect the sincerity of the ballot'. Violations of the law can result in a prison sentence of up to one year and a fine.

These French laws in some ways echo and in some ways differ from older attempts to control abusive communication and propaganda. For example, German law forbids dissemination and propagation of the symbols of "unconstitutional organizations" (*Strafgesetzbuch* §86). Historical cases in the US relate to public safety – it is not free speech to yell "fire" in a public cinema (see *Schenck v. United States*, 1919). Modern laws prohibiting the spread of fake news make similar trade-offs between free speech and public welfare. But modern laws also must contend what it means to punish a person outside their jurisdiction who can nonetheless harm their citizens. This problem also has historical precedent. The Prussian and French governments of the 1840s exiled Karl Marx, but his publications still reached and influenced the citizens of those nations. But the problem of social manipulation through technology practically automates this ability to influence another nation's citizens from abroad, and solutions at this scale bear little relation to what could be achieved in the 1840s.

2.4 Campaign Finance

One clear legal obligation in the US is that foreign nationals cannot direct, dictate, control, or directly or indirectly participate in any election-related activities [62]. Foreign nationals making financial contributions to campaigns, such as paying for staff or advertising, are an important, if not primary, example of such illegal behavior. Thus, existing campaign finance regulations help enforce the 'security policy' we are considering. Transparency requirements in campaign finance regulations are one method of getting online advertising platforms to limit propaganda and abuse. Facebook has cited advertiser spend transparency as one of its efforts to "help prevent foreign interference in elections" [1]. Such regulations are a positive driver of anti-abuse work. Anti-abuse technology and policy retains all the benefits and limitations discussed in Section 2.2. In this section, we will discuss the implications of campaign finance regulations on our networked world.

One thing to bear in mind is usability for those who are verifying the nationality of those buying advertising. Accounts will need to be verified via government-issued ID, presumably remotely. In 2018, Facebook used postcards with a security code on it as part of its system to verity residence within the US [49]. Secondly, there is a question of who actually bears the security cost of verified accounts. This requirement would plausibly increase the value of valid advertiser accounts of a specific nationality on the black market. Criminal specialization and resale is well-documented [29, 55]. The primary task of those myriad advertiser accounts is not to protect the integrity of national elections. Usable security teaches us would should not ask users to do much beyond their primary tasks, because they will avoid secondary tasks such as security if they reduce their productivity too much [5, 53].

While expecting online advertising companies to do basic know-your-customer due diligence may be feasible, if difficult, other aspects of tracking finances are more difficult in 2019 than even in 2009. Pseudonymous electronic currencies, such as bitcoin, make it easier for motivated individuals, let alone governments, to transfer money across national borders outside of regulations. This worry is not hypothetical. The Russians indicted for interfering in the 2016 US election allegedly "principally used bitcoin when purchasing servers, registering domains, and otherwise making payments in furtherance of hacking activity" [61, p. 21].

2.5 Education

Pedagogy has long considered education to be a bulwark against propaganda and information operations. For example, in 1947 a prominent researcher of mathematical pedagogy warned:

> "In the schools of Germany between 1933 and 1939 the operational techniques of mathematics were undoubtedly learned as effectively as anywhere else in the world, as the sons and daughters of passive parents were being prepared for 'responsible citizenship' in a Nazi culture. However, during these fateful years pregnant with disaster for the peoples of the earth, the teachers of the Third German Reich carried into their classrooms their warped and distorted ideas and through a process called "education" led their obedient students to accept the vicious proposition that they belonged to a superior race and were destined to rule the world. The ability to examine the quality of evidence supporting this proposition, to analyze the assumptions on which it is based and to use related understandings associated with the nature of proof was not developed through their study of mathematics nor would such outcomes be tolerated in a Nazi society. The skills and operations of mathematics may be its only contribution to responsible citizenship in an anti-democratic culture but much more

> is expected from the study of mathematics in the classrooms of a democracy" [19, p. 200].

Extensive case studies form the basis for this confidence in the positive effect of analytic thinking and rigorous argument via mathematics, rather than the mere "operational techniques" [18].

It would be tempting here to cite the litany of international rankings showing the US falling behind other countries in student test scores. But it is exactly this focus on test results and operational techniques that [19] warns against. The fact that public school students are evaluated based on a test of whether they can acquire the correct answers via the slated operational techniques is a failure of "the study of mathematics in the classrooms of a democracy." The sort of education needed to combat propaganda, in the 1930s as well as now, is in how to reason reliably. In the 1930s, [18] demonstrated such a curriculum successfully in public high school.

In the framing of modern behavioral economics literature, we are advocating a pedagogy that systematically attempts to overcome human cognitive biases. For practical intents and purposes, [18] presents such an approach. Our initial thought when composing this article was to wax poetic about how everyday primary school students should learn about intelligence analysis and cognitive biases, for example from an analysts' training textbook such as [33]. But the mathematics pedagogy literature beat us to that conclusion by 75 years.

In the UK, a Parliamentary report about combating 'disinformation' states "digital literacy should be a fourth pillar of education, alongside reading, writing, and maths" [15, p. 96]. Digital literacy may well be a good goal, and forms a helpful set of background skills and knowledge. However, the same can be said of digital literacy in 2019 as mathematics education in 1947—more is expected from this education than mere skills and operations, it must also include reasoning and analysis skills to make use of digital literacy.

In Ukraine, attempts to educate adults in media literacy resulted in mild increases in efforts to cross-check information [42]. The study does not attempt to identify whether this changed the respondent's mind, or how respondents reasoned through any cross-checking they did. The media literacy plan is being piloted with 8th and 9th graders to better effect [34]. The Ukrainian efforts seem aimed at inculcating a sense that propaganda is a thing, and that mass media may be biased. These facts are a minimum necessity and positive step. But they fall far short of the curriculum envisaged by [18].

Media literacy efforts are generally not aligned with what the usable security literature would suggest. Media literacy suggests that users take time away from their primary task to check certain features of communications or stories. While these may be helpful heuristics, they are not the user's primary task. This usable security perspective helps explain why media literacy efforts are not generally successful. This media literacy approach is quite different from the long-term education in good reasoning that [18] advocates. A usable security perspective would suggest that such a long-term effort, which realigns the users' skills—and motivation—to be in line with the policy of resisting foreign interference, is more likely to be successful.

3 OBSERVATIONS FROM A USABLE SECURITY PERSPECTIVE

Many of the examples above, such as deterring adversaries or knocking them offline, as was done to the IRA, are indirect. They have the potential to reduce the volume or intensity of information operations that individuals are exposed to, but they do not involve interaction with the 'user'. Such methods can be used to complement more user-facing interventions, which we look at in this section.

Many of the ideas from usable security can be expressed, in a very oversimplified manner, through the phrase "do not overburden the user," or, perhaps, "reduce the effort required from the user." It is easy to see how this applies to the problem of election interference: an average person, who views many articles, messages, websites, and other communication, does not have the time, expertise, or inclination to investigate whether or not each of them is an information operation. Many of the techniques discussed, both above and later in this section, can help in this regard—reducing the number of opportunities an adversary has to influence a user helps the user behave in a more secure manner; that is, they are less likely to be influenced. However, automated techniques can not—and probably should not, given the potential for censorship and loss of freedom—filter away all attempts at foreign election interference. People will have to make decisions about the legitimacy of information, about what can be trusted, and about what to believe.

Going beyond the simple message of 'reduce user effort,' usable security advises about how to get people to act securely. One important contribution is how to effectively achieve security behaviour change [3]. For example, as we will discuss below, education needs to do more than just provide information—it has to be both actionable and doable. Simply telling people to watch out for information operations and fake news will not be effective—most people will lack the necessary capability to identify such things, and will not have the motivation to learn or apply the skills needed. People must be willing to change their behaviour for education to be effective. This is a particularly difficult problem in this domain, as people's existing political beliefs may be part of a broader cultural identity that resists change. Furthermore, humans have a tendency to continue to believe what they already know, via cognitive biases such as anchoring and confirmation biases [13]. Some promising recent work [51] shows that a game that 'inoculates' its players against weak forms of fake news strategies can improve their detection ability.

However, even if education can be effective, people must first be aware that there is a problem and then care to spend time and effort learning and applying the skills needed to behave more securely [8]. Education campaigns can improve their chance for success if messages are targeted at and tailored for particular groups, if there are suitable services available to support the desired behaviour, if awareness messages come through many different channels, and if there are role models or champions demonstrating the desired behaviour; success can be hindered by making the information hard to understand or inaccessible, delivered to a general audience, or not consistent across sites [12]. It must be made easy for people to do the right thing [16].

Already we can see that there are some very big challenges. People read multiple sites, communication applications, and social networks; information operations can target them from any combination of their sources. Any effective solution will require coordination among these platforms, and consistent awareness education and easy-to-use tools to enable individuals to act securely. Based on the usable security literature, we advise that inconsistent advice and inaccessible or difficult-to-use tools will not succeed.

In the rest of this section we will look at some existing attempts at awareness communication, technological measures, and user-facing tools from a usable security perspective.

3.1 UK SHARE Checklist

The first intervention we examine is an attempt to educate. The UK government has recently (early April, 2019) released advice [27] for individuals to follow when considering whether or not to share stories and other content online. It is a checklist of five items:

- **Source**—Make sure that the story is written by a source you trust, with a reputation for accuracy. If it's from an unfamiliar organisation, check for a website's 'About' section to learn more.
- **Headline**—Always read beyond the headline. If it sounds unbelievable, it very well might be. Be wary if something doesn't seem to add up.
- **Analyse**—Make sure you check the facts. Just because you have seen a story several times, doesn't mean it's true. If you're not sure, look at fact checking websites and other reliable sources to double check.
- **Retouched**—Check whether the image looks like it has been or could have been manipulated. False news stories often contain retouched photos or re-edited clips. Sometimes they are authentic, but have been taken out of context.
- **Error**—Many false news stories have phony or look-alike URLs. Look out for misspellings, bad grammar or awkward layouts.

From a usable security perspective, there are issues with most of these suggestions. Generally, they fail to account for biases that affect human behavior, fail to account for the effort and skill required to follow the advice, and do not point to resources that people can use to fact-check or verify reputation.

People often trust sources that are known for spreading inaccurate information, and often believe that legitimate sources are biased or incorrect. People trust and share inaccurate or false stories from less reputable sources because they confirm previously held beliefs [30]. Seeing content that matches their beliefs may also reduce motivation to properly assess the credibility of the site, meaning less accurated heuristics are used [41]. It may also be easy to trust such sources because there is very little personal risk or possible harm involved [23]. Better advice would be to avoid sharing if you don't know the source at all, and to point to fact checking sites where the reputation and authenticity of a site can be verified.

Asking users to read beyond the headline—"if it sounds unbelievable, it very well might be"—is far too general, unspecific advice. It is similar to a previous UK phishing education campaign, discussed in [39], which ran with the slogan "if it sounds too good to be true, then it probably is." Both of these ignore the fact that people will engage with content (whether phishing site offering a good bargain or fake news story with attractive headline) that they respond to.

Advising users to analyze and check the facts is asking users to spend a lot of time and effort—which they will certainly retain for their primary tasks and not spend on the security task of verification [5] The advice also ignores the fact that people are biased to confirm their beliefs and disregard evidence that contradicts them. Countering a passionate, emotional stance by advocating rational decision-making and the expenditure of a lot of effort is not likely to be effective. Of course, this depends on context: when considering sharing on social media, it is particularly unlikely that users will be willing to spend time considering an action that likely just takes a click or two of the mouse; however, in other scenarios—emailing colleagues, for example—it might be more reasonable to expect users to be willing to invest time and effort checking facts. The advice to look at fact checking websites is useful—but again, no resources are provided.

A well-manipulated image is very hard to spot—skilled forensic scientists working for national crime agencies spend days and weeks on cases. Even advice [47] for detecting simple, unsophisticated manipulations would take a lot of time for each image. Furthermore, the ability to manipulate audio and video is improving rapidly; the development of 'deep fakes', which use neural networks to generate realistic audio and video footage has serious implications for the future [50].

Telling people to look for 'phony' URLs has been tried in phishing education, and has been found to be time-consuming and error-prone. Users are focused on their primary task—in this case, sharing information—and are unlikely to remain vigilant if they have to interrupt this task repeatedly. Furthermore, misspellings and bad grammar are not a reliable indicator; in the UK, 28% of adults have a standard of literacy of level 1 or below [36]. Many genuine communications will contain these errors, and so might be ignored if following this advice. Additionally, this advice is likely to be ineffective against many adversaries that have the resources to hire people with perfect language skills and set up realistic-looking domain names.

Overall, these five pieces of advice do not take any account of the time it would take to carry out these actions or the skills that would be needed to perform them correctly. It would add a huge effort tax on interacting with content. Herley, who looks at the cost-benefit trade-offs to users for following security advice [31], finds that it is rational for users to ignore many pieces of advice; in this case, where the costs of ignoring the device are not directly experienced by the individual, this is almost certainly true.

A more effective approach would involve showing how manipulation is done, why it is done, and the damage that out-of-control sharing of false or misleading information can do. Rather than assuming that all individuals are motivated by a desire to do good, effective education and advice must recognize that people have biases, seek to confirm what they already believe, and get a positive emotional response from sharing validating information with fellow believers. Getting individuals to understand the consequences

means they are more likely to take in interest in not being manipulated, which renders them amenable to changing their behavior—which simply presenting a list of advice that ignores their motivations will not do. This prepares the ground for them to learn specific skills [45].

3.2 Facebook

Facebook, as a platform, is a powerful tool for the communication of thoughts, ideas, and information. Whatever its benefits, it is also capable of being used to disseminate unsavory content and information operations. The report by the UK's House of Commons Digital, Culture, Media and Sport Committee [15], which looks at the role of 'disinformation' in recent elections, is highly critical of Facebook, citing attempts to mislead the committee about overseas interference in elections, and an unwillingness to be regulated or scrutinized.

Perhaps in response to the growing awareness of its use as a platform for information operations, Facebook has introduced features to try and increase transparency about who is sponsoring political content. They have introduced tools to show who is purchasing political advertising, and who is responsible for running pages. However, these features still require the user to be aware of them and then expend effort to assess the legitimacy of each page. The 'People who manage this page' feature states: "It's common for a Page to be managed by many people from different places. You can check for a mismatch between a Page's purpose and the location of the people who manage it." Like the SHARE checklist, this advice places the burden of effort onto the user, requiring them to first be aware of and find the tool, and then consider the purpose of the page and whether the locations of its administrators are consistent. Usable security teaches us that most users—even if they are aware that the information is present and that it should be used to check the authenticity of a page—will not bear this mental cost, and will not benefit from any added security.

Even if individuals use the tools Facebook has exposed, their effectiveness is unclear. The 'In the NOW' page, which states it "strives to build a community of mindful media consumers around important, curious and purpose-driven content" has 7 page managers from the United States, 2 from Germany, and 1 from Russia. Nothing seems inconsistent about the locations of the managers and the purpose of the page, and yet this was one of several pages suspended by Facebook for not declaring links to RT, the Russian state-operated media company [40]. The pages were eventually restored, and now carry a statement: "'In the Now" is a brand of Maffick which is owned and operated by Anissa Naouai and Ruptly GmbH, a subsidiary of RT.'

While paid political advertising and some state-funded pages have these tools—however effective they might be—there are also user-run groups and accounts without them that can conduct information operations. The DCMS report [15] states: "There also needs to be an acknowledgement of the role and power of unpaid campaigns and Facebook Groups that influence elections and referendums." Facebook removed a large number of accounts operated by the Internet Research Agency after the 2016 elections, removed 30,000 fake accounts ahead of the 2017 French elections [2], as well as a number of fake accounts and pages before the Moldovan EU elections [59].

From a usable security perspective, the detection and removal of fake accounts used to conduct information operations is a good solution, if Facebook can keep up. The detection needs to be comprehensive and adaptable; sophisticated adversaries can experiment, with little cost, to find different ways of avoiding detection, making this a challenging task. Detection and removal should happen as quickly as possible to limit the number of users exposed. Facebook does not have any stated public mitigation plan for what to do if a user is exposed to propaganda, as recommended in Section 2.3.

3.3 WhatsApp

WhatsApp is an end-to-end encrypted messaging application that allows users to send messages, photos, videos, and audio to other individuals or to groups. It became the vehicle for a large-scale election influence campaign in Brazil in 2018 [7]. Due to the high cost of internet access in the country, combined with free data to access WhatsApp, along with other applications, on mobile phone plans, the level of WhatsApp adoption is extremely high, and the ability of individuals to visit websites to verify or fact check any information they receive is limited. Information is shared and reshared between friends and family, which has the effect of increasing its credibility. When Bolsonaro supporters sent out message after message with misleading or manipulated information, photos, or videos, they rapidly went viral and spread around the population.

In response to this, as well as to incidents of mob violence in India incited by false information spread over the network, WhatsApp has added indicators to show when a message has been forwarded, and set a limit on how many times a message can be forwarded by an individual [4]. This drastically reduces the number of recipients one person can easily spread a message to.

Given the encrypted nature of the messaging platform, that is, the messages themselves cannot be observed by WhatsApp, this seems like a reasonable solution from a usable security perspective. The limit is unlikely to be detrimental to legitimate uses, but provides an additional barrier to the mass sharing of possibly false or unwanted information. As in other areas of cybersecurity, rate limiting increases the cost of attacks. Since the attack here requires repeated human interaction to forward, this should be effective, unless the requirement for human interaction can be subverted. Unfortunately, this solution is unique to private messaging apps. Social media companies designed for mass sharing to all of a user's connections cannot benefit from such rate limiting without changing their whole user model.

Another approach has been taken on the Line messenger app in Taiwan [56], which is also encrypted, like WhatsApp, so messages can not be observed directly. Users can forward messages to a bot run by a fact-checking service, which will reply with responses from fact checkers. The responses are saved in a database, so queries about a previously-seen message are handled instantly. With this solution, users must be aware of the service and motivated to use it; however for those who are aware, it seems to be an effective approach that does not require much effort from individuals. It also maintains users' autonomy: they have a choice

Should do	Should avoid
Help people understand consequences of their actions	Overloading the voters' primary task(s)
Align platform policy and communication with laws and regulations	Expecting voters to be IT and media analysis experts
Enable enforcement of existing laws via transparency	Ignoring voter cognitive biases
Long-term education in critical thinking	Educating voters as if we all can become news subject matter experts
Platforms should contain and reduce information operations technologically	Censorship – automated decision making needs to be transparent to non-experts

Table 1: An overview of our suggestions for a usable security approach to limiting foreign interference in US and UK elections via social manipulation through technology.

of whether or not to believe the fact checkers, and if multiple fact checkers have different opinions, the user is presented with all of them and can make their own choice.

3.4 Pros and cons of automated intervention

There are limits to what automated detection or other defenses can successfully achieve; relying on these too much raises questions about censorship and whether they interfere with democratic values. There are therefore limits to how much user effort can be minimized, and at some point individuals will be required to act or make decisions. A usable security perspective can help those defending elections to understand how to encourage more secure behaviour by the intended legitimate voters.

Twin concepts from the international relations literature help from the tension here. Therein, there are two notions of [physical] security: freedom from — negative security — and freedom to, or positive security. In this context, 'freedom to' tends to mean freedom to express certain shared values or actions. This conception frames "doing [positive] security [as] the identification and survival of the core values of the order" [48, p. 13]. One conception of the problem the US and UK are grappling with is to retain freedom from large-scale election interference while maintaining each citizens' freedom to express themselves. This tension mirrors the tension in the usable security literature between a security policy, which describes desired behaviour, and the users' goals.

One obvious drawback to automated interventions is that they may limit voter autonomy. Insofar as the interventions share and protect the goals and values expressed in the law, automated protection is a form of collective defense. However, the challenge, for which usable security can suggest solutions, is to get individuals to adopt and share these values and be motivated to act in their defense, which is necessary in cases beyond the ability or defined limits of automated protection.

4 CONCLUSIONS

Lessons from usable security tell us that important aspects of the effort to combat election interference via social manipulation are likely to fail. Table 1 summarizes what these lessons recommend responses should do and should avoid. In many instances, companies with misaligned incentives dodge work by pushing it on to the users. But we know that the users do not have the skill or incentive to perform the tasks asked of them, and that their compliance budget is likely to be exceeded should they attempt to.

One exception to our reticence to engage the user is long-term education, especially of young people. Combating propaganda has long been a focus of educational efforts. As the problem is likely an enduring one, the social cost of teaching all children mathematical reasoning skills and digital literacy is a plausible solution, though it will take a long time scale to enact. The usable security literature has lessons for how to make this education more successful, as well. Do not exhort students with abstract needs, as the SHARE [27] model does. Provide concrete examples of the consequences of sharing propaganda and fake news. When coupled with the mathematics pedagogy literature, this could make a powerfully engaging curriculum.

Attempts to educate older users to behave more responsibly when sharing content online must start by acknowledging that people have biases and want to see information that supports their beliefs, and get a positive emotional response from sharing that information with fellow believers. Assuming that all users are fair, and unbiased, and want to do good—and just need to be given a list of instructions how to do it—ignores this and will not be effective. Instead, the strategy has to make people aware that there are consequences of their behavior that they do not want. This awareness can lead to changes in behavior to avoid the undesired consequences.

Outside of education, solutions that do not burden the user are an important tool. This recommendation is equally true of technical and political solutions. For example, WhatsApp's rate limiting of message forwarding appears to be an elegant solution to their propaganda spam problem. To the extent possible, platforms should be built with such features in mind before they're exploited, not reactively. To the extent possible, governments should re-align the incentives of the technology platforms to prevent the platforms from pushing this important national-security work on users. Of course, what work the platform does on behalf of the users should be transparent in an way that is explainable to even marginalized or at-risk users.

Election interference via social manipulation remains a difficult problem. In many ways, it predates the Internet. There are highly motivated adversaries and complex interlocking constraints on defenders. The discipline of usable security has useful insights to add to the many voices in this debate. No one discipline has all the solutions to election interference, but the lens of usable security is a useful one for helping design solutions that will work as intended in the real world.

4.1 Future Work

There are many things we have mentioned—often quite briefly—here that pose great challenges for future work. For example, we

have discussed how there is a trade-off between the amount of work automated systems can perform to limit election interference and human freedom and autonomy. A heavy-handed approach to automation can lead to censorship; but what is not automated will rely on human effort and ability. What is the correct balance? Where should an algorithm draw the line on what is fake and what is legitimate—and who decides which is which? How much can people be expected to do? Another important area is how to effect behaviour change. What is the best way to teach people the importance of and skills needed to make more secure decisions? How can this message be coordinated among different web platforms? How can these tasks be supported by tools? These are all open—and difficult—challenges, that will need to be addressed by researchers as well as society in order to make progress towards tackling this problem.

ACKNOWLEDGEMENTS

The authors thank Michaela Stone for pointing us to Fawcett's work on mathematics pedagogy.

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

References herein to any specific commercial product, process, or service by trade name, trade mark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by Carnegie Mellon University or its Software Engineering Institute.

[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution. DM19-0535

REFERENCES

[1] Richard Allan. 2019. Protecting Elections in the EU. https://newsroom.fb.com/news/2019/03/ads-transparency-in-the-eu/. Accessed 2019-06-30.

[2] Eric Auchard and Joseph Menn. 2017. Facebook cracks down on 30,000 fake accounts in France. https://www.reuters.com/article/us-france-security-facebook/facebook-cracks-down-on-30000-fake-accounts-in-france-idUSKBN17F25G. Accessed 2019-04-18.

[3] Maria Bada, Angela M Sasse, and Jason RC Nurse. 2019. Cyber security awareness campaigns: Why do they fail to change behaviour? *arXiv preprint arXiv:1901.02672* (2019).

[4] BBC News. 2019. WhatsApp restricts message-sharing to fight fake news. https://www.bbc.co.uk/news/technology-46945642. Accessed 2019-04-18.

[5] Adam Beautement, M Angela Sasse, and Mike Wonham. 2009. The compliance budget: managing security behaviour in organisations. In *New Security Paradigms Workshop*. ACM, 47–58.

[6] Emily Bell. 2018. Facebook creates Orwellian headace as news is labelled politics. *The Guardian* (24 Jun 2018). https://www.theguardian.com/media/media-blog/2018/jun/24/facebook-journalism-publishers

[7] Luca Belli. 2018. WhatsApp skewed Brazilian election, proving social media's danger to democracy. https://theconversation.com/whatsapp-skewed-brazilian-election-proving-social-medias-danger-to-democracy-106476. Accessed 2019-04-18.

[8] M Beyer, S Ahmed, K Doerlemann, S Arnell, S Parkin, A Sasse, and N Passingham. 2016. Awareness is only the first step: A framework for progressive engagement of staff in cyber security. *Business white paper: Hewlett Packard* (2016).

[9] Matt Bishop and David Wagner. 2007. Risks of e-voting. *Commun. ACM* 50, 11 (2007), 120–120.

[10] Samantha Bradshaw and Philip N Howard. 2018. Challenging truth and trust: A global inventory of organized social media manipulation. *The Computational Propaganda Project* (2018).

[11] Finn Brunton. 2013. *Spam: a shadow history of the Internet.* MIT Press, Cambridge, MA.

[12] Lynne Coventry, Pamela Briggs, John Blythe, and Minh Tran. 2014. Using behavioural insights to improve the public's use of cyber security best practices. *Gov. UK report* (2014).

[13] Pat Croskerry, Geeta Singhal, and Sílvia Mamede. 2013. Cognitive debiasing 1: Origins of bias and theory of debiasing. *BMJ Qual Saf* 22, Suppl 2 (2013), ii58–ii64.

[14] Anthony Cuthbertson. 2019. Russia plans to briefly disconnect from the internet to see what happens. https://www.independent.co.uk/life-style/gadgets-and-tech/news/russia-internet-disconnect-experiment-isp-cyber-war-putin-data-a8773961.html. Accessed 2019-04-18.

[15] Digital, Culture, Media and Sport Committee. 2019. *Disinformation and 'fake news': Final Report.* Technical Report. United Kingdom House of Commons.

[16] ENISA. 2018. Cybersecurity Culture Guidelines: Behavioural Aspects of Cybersecurity. https://www.enisa.europa.eu/publications/cybersecurity-culture-guidelines-behavioural-aspects-of-cybersecurity.

[17] Henry Farrell and Bruce Schneier. 2018. *Common-Knowledge Attacks on Democracy.* Technical Report 2018-7. Berkman Klein Center, Cambridge, MA, USA.

[18] Harold P Fawcett. 1938. *The nature of proof.* Number ED 096 174. National Council of Teachers of Mathematics, Reston, VA, USA.

[19] Harold P Fawcett. 1947. Mathematics for responsible citizenship. *The Mathematics Teacher* 40, 5 (1947), 199–205.

[20] Ariel J. Feldman, J. Alex Halderman, and Edward W. Felten. 2007. Security Analysis of the Diebold AccuVote-TS Voting Machine. In *Workshop on Accurate Electronic Voting Technology (EVT'07)*. USENIX Association.

[21] Martin Flintham, Christian Karner, Khaled Bachour, Helen Creswick, Neha Gupta, and Stuart Moran. 2018. Falling for fake news: investigating the consumption of news via social media. In *Human Factors in Computing Systems (CHI)*. ACM, Montreal, QC, Canada.

[22] Agence France-Presse. 2018. Indian state cuts internet after lynchings over online rumours. https://www.theguardian.com/world/2018/jun/29/indian-state-cuts-internet-after-lynchings-over-online-rumours. Accessed 2019-07-03.

[23] Batya Friedman, Peter H. Khan, Jr., and Daniel C. Howe. 2000. Trust Online. *Commun. ACM* 43, 12 (Dec. 2000), 34–40. https://doi.org/10.1145/355112.355120

[24] Daniel Funke and Daniela Flamini. [n. d.]. A guide to anti-misinformation actions around the world. https://www.poynter.org/ifcn/anti-misinformation-actions/.

[25] Vijaya Gadde and Yoel Roth. 2018. Enabling further research of information operations on Twitter. https://blog.twitter.com/official/en_us/topics/company/2018/enabling-further-research-of-information-operations-on-twitter.html. Accessed 2019-04-12.

[26] Axel Gelfert. 2018. Fake News: A Definition. *Informal Logic* 38, 1 (2018), 84–117. https://doi.org/10.22329/il.v38i1.5068

[27] HM Government. 2019. Don't feed the beast. https://sharechecklist.gov.uk/. Accessed 2019-04-18.

[28] Andy Greenberg. 2019. US Hackers' Strike on Russian Trolls Sends a Message—but What Kind? https://www.wired.com/story/cyber-command-ira-strike-sends-signal/. Accessed 2019-04-18.

[29] Chris Grier, Lucas Ballard, Juan Caballero, Neha Chachra, Christian J. Dietrich, Kirill Levchenko, Panayiotis Mavrommatis, Damon McCoy, Antonio Nappa, Andreas Pitsillidis, Niels Provos, M. Zubair Rafique, Moheeb Abu Rajab, Christian Rossow, Kurt Thomas, Vern Paxson, Stefan Savage, and Geoffrey M. Voelker. 2012. Manufacturing Compromise: The Emergence of Exploit-as-a-service. In *ACM Conference on Computer and Communications Security (CCS '12)*. Raleigh, North Carolina, USA, 821–832.

[30] Craig A Harper and Thom Baguley. 2019. "You are Fake News": Ideological (A)symmetries in Perceptions of Media Legitimacy. https://doi.org/10.31234/osf.io/ym6t5

[31] Cormac Herley. 2009. So Long, and No Thanks for the Externalities: The Rational Rejection of Security Advice by Users. In *Proceedings of the 2009 Workshop on New Security Paradigms Workshop (NSPW '09)*. ACM, New York, NY, USA, 133–144. https://doi.org/10.1145/1719030.1719050

[32] Edward S Herman and Noam Chomsky. 1988. *Manufacturing Consent* (2nd ed.). Pantheon Books, New York.

[33] Richards J Heuer, Jr. 1999. *Psychology of intelligence analysis.* US Central Intelligence Agency.

[34] Sasha Ingber. 2019. Students In Ukraine Learn How To Spot Fake Stories, Propaganda And Hate Speech. *National Public Radio* (22 Mar 2019). https://www.npr.org/2019/03/22/705809811/students-in-ukraine-learn-how-to-spot-fake-stories-propaganda-and-hate-speech

[35] Karen E Jacowitz and Daniel Kahneman. 1995. Measures of anchoring in estimation tasks. *Personality and Social Psychology Bulletin* 21, 11 (1995), 1161–1166.

[36] Sophie Jamieson. 2016. 'Three Rs' on the decline as a quarter of adults have a reading age so low they struggle to read a bus timetable. https://www.telegraph.co.uk/news/2016/08/28/three-rs-on-the-decline-as-a-quarter-of-adults-have-a-reading-ag/. Accessed 2019-04-18.

[37] Mohammad Hanif Jhaveri, Orcun Cetin, Carlos Gañán, Tyler Moore, and Michel Van Eeten. 2017. Abuse reporting and the fight against cybercrime. *Computing Surveys (CSUR)* 49, 4 (2017), 68.

[38] Joint Chiefs of Staff. 2014. *Information Operations.* Technical Report JP 3-13. U.S. Dept of Defense, Washington, D.C.

[39] I. Kirlappos and M. A. Sasse. 2012. Security Education against Phishing: A Modest Proposal for a Major Rethink. *IEEE Security Privacy* 10, 2 (March 2012), 24–32. https://doi.org/10.1109/MSP.2011.179

[40] Tom McKay. 2019. Facebook Suspends Three Pages With Millions of Video Views, Saying They Need to Disclose Russia Ties. https://gizmodo.com/facebook-suspends-three-pages-with-millions-of-video-vi-1832679030. Accessed 2019-04-18.

[41] Miriam J. Metzger, Andrew J. Flanagin, and Ryan B. Medders. 2010. Social and Heuristic Approaches to Credibility Evaluation Online. *Journal of Communication* 60, 3 (9 2010), 413–439. https://doi.org/10.1111/j.1460-2466.2010.01488.x

[42] Erin Murrock, Joy Amulya, Mehri Druckman, and Tetiana Liubyva. 2017. *Winning the war on state-sponsored propaganda.* Technical Report. International Research and Exchanges Board, Washington, DC.

[43] Lawrence D Norden and Christopher Famighetti. 2015. *America's Voting Machines at Risk.* Brennan Center for Justice at New York University School of Law, New York, NY, USA.

[44] George Orwell. 1949. *Nineteen Eighty-Four: A Novel.* Secker & Warburg, London.

[45] Simon Parkin, Elissa M Redmiles, Lynne Coventry, and M Angela Sasse. 2019. Security When it is Welcome: Exploring Device Purchase as an Opportune Moment for Security Behavior Change. In *Workshop on Usable Security.*

[46] Gordon Pennycook, Adam Bear, Evan Collins, and David G. Rand. 2019. The implied truth effect: Attaching warnings to a subset of fake news stories increases perceived accuracy of stories without warnings. *SSRN pre-print* (mar 2019). https://ssrn.com/abstract=3035384

[47] Alicia Prince. [n. d.]. This is How You Can Tell if an Image has Been Photoshopped. https://www.lifehack.org/articles/technology/this-how-you-can-tell-image-has-been-photoshopped.html/. Accessed 2019-04-18.

[48] Paul Roe. 2008. The 'value'of positive security. *Review of international studies* 34, 4 (2008), 777–794.

[49] Vanessa Romo. 2018. Facebook is counting on postcards to prevent future election interference. *National Public Radio* (20 Feb 2018). https://www.npr.org/sections/thetwo-way/2018/02/20/587070994/facebook-is-counting-on-postcards-to-prevent-future-election-interference

[50] Dave Roos. 2017. Manipulated Video and Audio Will Make Future Fake News Even More Believable. https://www.seeker.com/manipulated-video-and-audio-will-make-future-fake-news-even-more-belie-2181890359.html. Accessed 2019-04-18.

[51] Jon Roozenbeek and Sander van der Linden. 2019. Fake news game confers psychological resistance against online misinformation. *Palgrave Communications* 5, 65 (2019).

[52] Russian Federation. 1993. Constituion of the Russian Federation. https://web.archive.org/web/20101016230423/http://archive.kremlin.ru/eng/articles/ConstEng2.shtml.

[53] M. A. Sasse, S. Brostoff, and D. Weirich. 2001. Transforming the 'Weakest Link' — a Human/Computer Interaction Approach to Usable and Effective Security. *BT Technology Journal* 19, 3 (01 Jul 2001), 122–131. https://doi.org/10.1023/A:1011902718709

[54] R. Shirey. 2007. Internet Security Glossary, Version 2. RFC 4949 (Informational). , 365 pages.

[55] Aditya K Sood and Richard J Enbody. 2013. Crimeware-as-a-service: A survey of commoditized crimeware in the underground market. *International Journal of Critical Infrastructure Protection* 6, 1 (2013), 28–38.

[56] Kirsten Han Splice. [n. d.]. Taiwanese Cofacts fact-checks information on LINE. https://ijnet.org/en/story/taiwanese-cofacts-fact-checks-information-line.

[57] Jonathan M Spring. 2013. Modeling malicious domain name take-down dynamics: Why eCrime pays. In *eCrime Researchers Summit (eCRS).* IEEE, San Francisco.

[58] Jonathan M Spring, Sarah Kern, and Alec Summers. 2015. Global adversarial capability modeling. In *APWG Symposium on Electronic Crime Research (eCrime).* IEEE, Barcelona.

[59] Alexander Tanas and Alissa de Carbonnel. 2019. Ahead of EU polls, Facebook voids accounts targeting Moldovan election. https://www.reuters.com/article/us-facebook-moldova/ahead-of-eu-polls-facebook-voids-accounts-targeting-moldovan-election-idUSKCN1Q312B. Accessed 2019-04-18.

[60] Joshua A Tucker, Andrew Guess, Pablo Barberá, Cristian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal, and Brendan Nyhan. 2018. *Social media, political polarization, and political disinformation: A review of the scientific literature.* Technical Report. William and Flora Hewlett Foundation.

[61] United States District Court for the District of Columbia. 2018. *United States of America v. Netyksho et al.* Number 1:18-cr-00215-ABJ. https://www.justice.gov/file/1080281/download

[62] U.S. Federal Election Commission. 2017. Foreign nationals. https://www.fec.gov/updates/foreign-nationals/. Accessed 2019-03-18.

[63] Samuel C Woolley. 2016. Automating power: Social bot interference in global politics. *First Monday* 21, 4 (2016).

[64] Savvas Zannettou, Tristan Caulfield, Jeremy Blackburn, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Guillermo Suarez-Tangil. 2018. On the Origins of Memes by Means of Fringe Web Communities. In *Proceedings of the Internet Measurement Conference 2018 (IMC '18).* ACM, New York, NY, USA, 188–202. https://doi.org/10.1145/3278532.3278550

[65] Savvas Zannettou, Tristan Caulfield, William Setzer, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. 2019. Who Let The Trolls Out?: Towards Understanding State-Sponsored Trolls. In *Proceedings of the 10th ACM Conference on Web Science.* ACM, 353–362.

[66] Jiang Zemin. 2000. Extradition Law of the People's Republic of China (Order of the President No. 42). https://www.unodc.org/res/cld/document/chn/2000/extradition_law_of_the_peoples_republic_of_china_html/China_Extradition_Law_2000.pdf.