Cyber Intelligence and Influence: In Defense of "Cyber Manipulation Operations" to Parry Atrocities

*Rhiannon Neilsen**

If you can spy on a network, you can manipulate it. . .
The only thing you need is active will.
—Michael Hayden
Former director, U.S. National Security Agency
Former director, Central Intelligence Agency[1]

Eddie Chapman was a scallywag and scoundrel, notorious for his frequent run-ins with the law for lying, cheating, and stealing in the 1930s and 1940s. But during World War II, he was also "Agent Zigzag"—a double agent working for rival intelligence agencies, the Nazi Abwehr and the U.K.'s MI5. In 1944, Chapman was tasked by the Nazi regime to report on the success of their V-1 and V-2 rockets targeting London. Zigzag—operating at the directive of MI5—consistently falsified the results of the rockets so that the Nazi leadership would alter their targets to (unbeknownst to them) hit less populated areas of London. This ultimately resulted in fewer civilian deaths.[2]

As the above vignette highlights, spies have long been understood to not just obtain information but to interfere with or obstruct the operations to which they are privy. This includes spreading disinformation about capabilities or

Rhiannon Neilsen, Stanford University, Stanford, California (neilsen@stanford.edu)

*I am very grateful to Cécile Fabre, Juan Espindola, Ross Bellaby, Alex Leveringhaus, Ron Dudai, Toni Erskine, Herb Lin, Janina Dill, Lachlan Shelley, Corinne Dale, and Paul John for their incisive feedback on earlier iterations of this essay and its arguments. My thanks also go to the Stanford University Political Theory and Technology Group (Rob Reich, Henrik Kugelberg, Diana Acosta Navas, and Sheon Han) and to the editors of *Ethics & International Affairs* for helping improve the piece.

Ethics & International Affairs, 37, no. 2 (2023), pp. 161–176.

doi:10.1017/S0892679423000187

plans, withholding information, falsifying reports, and planting fabricated information to be intentionally picked up and used by the enemy. Perhaps the most famous example of this is Operation Fortitude, where (among other deception tactics) the Allies broadcast fake radio chatter intended to deceive the Nazis into believing the D-Day landing would take place near Pas-de-Calais, not Normandy.[3] According to Sun Tzu, "We can cause the doomed spy to carry false tidings to the enemy."[4] It seems Sun Tzu was interested in using "doomed spies" to disseminate fake versions of *one's own* plans or capacities to mislead the enemy, as in Operation Fortitude.[5]

In this essay, I am concerned instead with the ethics of covertly manipulating the *enemy's* intelligence—similar to Agent Zigzag. I focus on the use of distinctly twenty-first century cyberspace capabilities to plant false information or fabricate content that will reside in the adversary's systems.[6] Specifically, I ask: Is a cyber spy or cyber traitor ethically permitted to manipulate, falsify, or plant misleading information in the adversary's networks (via cyber means) to prevent impermissible acts—notably genocides, war crimes, crimes against humanity, and ethnic cleansing?[7] Are cyber spies *obligated* to manipulate or falsify such information/orders rather than simply steal that information? To answer these questions, I tease out the distinction between "cyber espionage"—canvassed by Cécile Fabre as part of "cyber intelligence" (CYBINT)—and what might be regarded as "cyber manipulation." I then consider the ethics of using (what I call) "cyber manipulation operations" (CMOs).

As part of my focus on extraordinary situations in which atrocities occur, there are a few caveats to what I am suggesting. First, I am concerned only with spies and traitors that are legitimately acting on behalf of government intelligence agencies that have ordered them to engage in CMOs. I raise the question of permissible vigilantism in the conclusion. Second, I begin from the premise that CMOs must be considered once traditional efforts (for instance, publicly appealing to respect for human rights, bargaining with the target regime, pursuing diplomatic channels, and "naming and shaming") no longer appear to have a reasonable prospect of success alone (but must continue nonetheless).[8] As per Sissela Bok's requirement: "In any situation where a lie is a possible choice, one must first seek truthful alternatives."[9] It is precisely when honest attempts to dissuade perpetrators from committing atrocities fail (as they so often do) that we are confronted with the question of whether secretly manipulating an adversary's intelligence (as it pertains to the atrocities) might be justified. Third, and following Fabre, I accept

https://doi.org/10.1017/S0892679423000187 Published online by Cambridge University Press

the prevailing view that deception, lying, and intentional manipulation are pro tanto impermissible.[10] This is because, as Fabre explains, deception abuses and subverts trust, curtails a person's agency to act freely, gets the person to "act as one wishes without their informed consent, thereby treating her as a means to one's ends," and frustrates people's capacity to communicate and work collaboratively as equal "moral and rational agents," ultimately hindering the "prospects of a flourishing life."[11] Yet I also agree with her that there are special circumstances where such measures may be permissible or even required.[12] This essay thus seeks to determine whether deceiving potential perpetrators via cyber manipulation to prevent atrocity crimes is one such case.

The essay proceeds as follows. In the first section, I briefly recount Fabre's notions of cyber espionage and "cyber sabotage." In the second section, I describe how CMOs might help prevent atrocities and make the prima facie case for their permissibility. The third section considers—and responds to—concerns about using CMOs. I conclude the essay in section four.

Cyber Espionage

Writing on the ethics of espionage, Fabre notes that spies are often asked to do wrongful acts in our name, "on our behalf and at our behest."[13] She argues that spies are pro tanto permitted and indeed obligated (in some cases) to do so if the purpose is to forestall unjust acts. As part of her analysis, Fabre examines the ethics of using cyber intelligence to observe, gather, or steal information (including about human sources and other agents) via cyberspace.[14] It is assumed that such CYBINT operations are permissible if they will (in part) thwart rights transgressions; that is, if the intelligence collected will help agents to fulfill their duty to avert an unjust act. According to Fabre, conducting espionage (cyber or otherwise)—which involves tactics of manipulation and deception—to gain information that would avert foundational rights violations can be permissible, if not mandatory, so long as it is necessary, proportionate, and effective.[15] Cyber espionage may take the form of collecting, stealing, and releasing information so that states (specifically, governments) are better placed to act to parry atrocities, for instance.[16] After all, it is generally accepted that states and the international community (through the United Nations) bear moral duties to protect populations from genocide, war crimes, crimes against humanity, and ethnic cleansings. This much is perhaps best evidenced by the Responsibility to

https://doi.org/10.1017/S0892679423000187 Published online by Cambridge University Press

Protect (RtoP) norm, which was universally agreed to by all United Nations states at the 2005 United Nations World Summit.[17] I agree with Fabre that it is sometimes ethically acceptable (and, in some cases, required) to use cyber espionage to clandestinely steal information pertaining to a regime's or a nonstate armed group's ability to execute atrocities.[18] Consider the following original hypothetical case concerning extermination camps ("Extermination Camp" scenario):

> State *X* has a suspicion that State *Y* is building extermination camps that will be used to systematically slaughter a certain long-oppressed minority group (*M*) in State *Y*. Publicly available online and offline hate speech campaigns conducted by State *Y* regarding *M* has sharply increased, and now State *X*'s surveillance satellites find what appears to be the initial stages of an extermination camp being built.

Due to the online hate speech and satellite imagery, State *X* reasonably believes that State *Y* is planning to conduct grievous rights violations. Following Fabre's requirement for evidenced-based reasoning, State *X* would be pro tanto permitted to use cyber capabilities to hack into State *Y*'s governmental computer databases to steal information to verify its concerns.[19]

However, it is often not the lack of evidence that is the issue: it is motivating states to "do something" (although not just anything) to prevent and mitigate such crimes that is the greater obstacle. My attention, therefore, is concerned with other, covert cyber operations that likely (but not necessarily) take place *after* the justified cyber espionage.[20] One option that Fabre briefly mentions is cyber sabotage, which she discusses in the context of the 2010 Stuxnet cyber worm (which rendered Iranian nuclear centrifuges inoperable) and the (Russian-launched) distributed denial-of-service, or DDoS, attacks on Estonia in 2008. Yet, she claims that these cyber operations "take us into the territory of cyberwarfare" and are outside her focus on cyber intelligence.[21] I agree. These cyberattacks aimed to severely "disrupt, degrade, diminish, or destroy"[22] the very *cyber systems* they are targeting. This is qualitatively different from what I am suggesting: as I discuss below, I am talking about cyber operations that change the *information resident within* the systems and thus influence the *human operators* expected to act based on that information. In this sense, I contend that there are cyber operations that sit closer to cyber espionage, conducted as part of CYBINT, than to "cyberattacks" or "cyber warfare"; namely, cyber manipulation. I explain more in the next section.

https://doi.org/10.1017/S0892679423000187 Published online by Cambridge University Press

Cyber Manipulation

A range of behaviors fall under the umbrella of "manipulation." For instance, stealing data may manipulate a target's behavior: by obtaining such sensitive information, we equip ourselves with information that allows us to act in a particular way, and the enemy will respond to our revised plan in kind. Openly attacking networks via cyberspace, or threatening to do so, may also effectively manipulate perpetrators into refraining from committing crimes, thus acting as a form of deterrence.[23] By "cyber manipulation operations," I mean covertly hacking into an adversary's networks and (in addition to collecting data) altering or falsifying the information resident in the enemy's systems with the aim of furtively *influencing the autonomy of the human agent* by intentionally misleading her—à la "Agent Zigzag 2.0." This could include clandestinely falsifying orders (to send militants to the wrong place, for instance), supply chains (ordering the wrong kinds of ammunition, for example), the locations of victims, the outcomes of missions, military capabilities, and blueprints resident in the perpetrators' networks.

In this way, CMOs are distinct from directly coercing, blackmailing, or extorting a human operative via cyberspace into being a double agent.[24] With CMOs, there is no requirement to establish and maintain contact with a human target, long term or otherwise, as per traditional human intelligence (HUMINT) operations. Rather, it is the information in the cyber systems that is manipulated to present a falsified image or message to the human user as part of CYBINT/cyber espionage.[25] The purpose of CMOs is to get the human targets to act—or not act—based on their own (now falsified) data, which they have no reason to believe would be false.

Cyber manipulation is further dissimilar to what Fabre means by cyber *sabotage*. In the former, it is the content resident in the adversary's systems that is being manipulated to influence the behavior of the human consumers of that content. The latter, by contrast, is geared toward damaging the cyber systems themselves. CMOs include hacking into networks and then distorting the existing data within those systems; furtively tampering with or falsifying orders from chains of command; manipulating the logistical lines required for mass killing; adjusting capabilities of military equipment; altering the reports of missions; misdirecting troops; and sending emails (from an authority's account) to change or delay plans or to "stand down" entirely.

https://doi.org/10.1017/S0892679423000187 Published online by Cambridge University Press

An example of hacking into an adversary's networks and sending orders from that network occurred immediately prior to the U.S. invasion of Iraq in 2003. The United States hacked into the internal network of Iraq's Ministry of Defence and sent emails from that network encouraging desertion from then president, Saddam Hussein. According to Barbara Starr, "The disguised emails, being sent to key Iraqi leaders [from within the Iraqi network], urged them to give up, to dissent and to defect. If they do not, the messages warn, the United States will go to war with them."[26] The emails aimed to "convince the Iraqi leadership they cannot win a war against the United States and its allies" and the "US military and intelligence officials [hoped] that the Iraqis [did] not realize where the e-mails [were] coming from."[27] Moreover, the email also apparently included instructions for those Iraqi military leaders on how to defect by contacting the United Nations in Iraq.[28] Reconstructing what might have been read "by, say, an Iraqi Army brigadier general in charge of an armoured unit outside of Basra," Richard Clarke and Robert Knake surmise that the email would have said something like: the U.S. military will "overwhelm forces that oppose" them, that the United States "do[es] not want to harm you or your troops," and suggested that Iraqi troops "walk away . . . go home."[29] According to Clarke and Knake, the messages were successful—a considerable number of officers took the suggestion: units "neatly" left their tanks outside their bases, commanders ordered their subordinates home hours prior to the invasion, and the troops donned civilian clothes and tried to leave.[30]

Twenty years later, in March 2023, hacktivists from a Ukrainian "Cyber Resistance" group (which has ties to the Ukrainian government) sent false emails from within a secure network amid Russia's unjust invasion of Ukraine.[31] The hackers (in addition to collating evidence in the email inbox) sent fabricated emails to the wife of Russian Airforce colonel, Atroshchenko Sergey Valeriyovych—a man accused of committing war crimes, including ordering the killing of six hundred civilians in Mariupol.[32] As part of this operation, the hackers successfully convinced the colonel's wife to organize a "Patriotic Photoshoot," which helped reveal the identities of the Russian pilots in Atroshchenko's 960th Assault Aviation Regiment.[33] While this was not a CMO that aimed to curtail atrocities specifically, it further highlights the potentiality of secretly influencing the enemy via cyberspace.

Additional covert CMOs may also include distorting or fabricating information within the target's cyber systems and official networks so that it appears to be coming from the atrocity-perpetrating *military leaders themselves* (ideally, so that such interference goes undetected for a time). Following from the example

https://doi.org/10.1017/S0892679423000187 Published online by Cambridge University Press

above, such emails could have been manipulated to appear as though they had come from Hussein's *top generals* (not just from within the secure network). Consider the following hypothetical situation concerning emails ("Email" scenario), building on the "Extermination Camp" scenario previously described:

> Hackers from State *X* have successfully penetrated the computer networks of State *Y*. The hackers can compose and send (unbeknownst to the owner of the email) slightly altered or contradictory orders to the owner's subordinates, instructing them to delay, adjust, or relocate. These orders are not radical enough to raise suspicion, but are sufficient to redirect, forestall, or confuse those subordinates to the extent that they slow, postpone, or cease their preparations for atrocities.

In addition to sending emails, it is also possible to manipulate communications in more subtle ways. For instance, according to Captain (now Major) Stephen Whitham, a computer scientist in the Army Cyber Institute at West Point: It is "technically possible to fool a sender into thinking a message has been sent, while preventing the receiver from ever receiving that message."[34] Data could be "misrouted" by "deliberately changing destination headers of Internet packets."[35] To this extent, CMOs can be used to "withhold important messages (like orders or emails)" from potential perpetrators.[36] A man-in-the-middle attack could also be used, whereby cyber spies could ensure that "only information that the [cyber]attacker allows will then pass from or to the victim . . . so the victim gets or sends all their normal volume of data but no one can read it."[37]

We can also modify the "Extermination Camp" scenario, in the form of a hypothetical situation involving the blueprints for the extermination camp ("Blueprints for the Extermination Camp" scenario), to highlight another way in which a foreign intelligence agent or agency could deploy a CMO.

> After hacking into State *Y*'s networks and discovering evidence of an extermination camp, State *X*'s cyber spies falsify the construction site's blueprints, logistical production lines, resource orders, and supply chains. The hackers also surreptitiously cancel, delay, and redivert certain materials required for the construction of the extermination camps. This serves to prolong or complicate the construction process, and thus the onset of the killing. All the while, the cyber spies create a "feedback" loop to evade detection, wherein those monitoring the network do not detect any false activity.[38]

The acts conducted here as part of the CMOs are pro tanto ethically permissible on four counts. First, engaging in such activities is likely to be more *effective* for human protection than (for instance) merely stealing and (perhaps) publicly

https://doi.org/10.1017/S0892679423000187 Published online by Cambridge University Press

releasing State *Y*'s plans for extermination camps or deleting such plans altogether. This is because deleting or stealing and publicly releasing the adversary's information would invariably raise the alarm within the genocidal regime that their systems have been breached. In response, State *Y* would likely set about identifying and patching the network vulnerability, thereby increasing its cyber defensive infrastructures. Any future cyber espionage, surveillance, or disruptive cyber operations conducted by State *X* via the same access point in the network would thus be foreclosed. If it would be effective, proportionate, and necessary for State *X*'s cyber spies to manipulate or plant false information in State *Y*'s system *in addition to* or *instead of* merely stealing, collating, or disclosing information pertaining to atrocities, State *X* has an obligation to do so.

The permissibility or requirement of waging cyber manipulation operations does not hinge on the promise to be effective at *entirely* halting atrocities, although this would be ideal. CMOs may be permissible, even mandatory, if it allows states, acting in accordance with RtoP, to "buy time" to (for instance) form *better* atrocity prevention plans (rather than knee-jerk responses). It may also buy time for states to form what Toni Erskine calls a "coalition of the obligated" to intervene (proportionality, necessity, and effectiveness considered);[39] intensify diplomatic negotiations with the target state; or (at the very least) mitigate the severity and extent of the killing. States can increase political conversations with the target state while continuing to surreptitiously subvert the genocidal regime's plans; this is assuming the CMO continues to go undetected. By the same token, such CMOs may also afford genocidal regimes time to "think twice" and abandon their pursuit of unjust policies. Relatedly, the use of CMOs does not ipso facto preclude State *X* from engaging in other protection efforts. Of course, State *X* continues to bear a responsibility to detect mass atrocities (including via CYBINT) and protect vulnerable populations in State *Y* beyond the use of cyber manipulation. The intention here is merely to highlight that State *X* may permissibly consider conducting CMOs as one way to satisfy protection duties.

Second, altering orders, fabricating or amending plans, or impersonating a figure of authority via CMOs are predominantly *discriminate* forms of cyber deception. The forged content (as in the "Email" and "Blueprints for the Extermination Camp" scenarios) will be read only by those who are already privy to, and thus likely to be *complicit in*, the atrocity plans. So, following Fabre's argument, the targets are liable to being duped in this way.[40]

https://doi.org/10.1017/S0892679423000187 Published online by Cambridge University Press

Third, CMOs promise to be comparatively *less harmful* than other protective efforts, such as widespread economic sanctions or armed humanitarian interventions. According to the jus in bello principle of necessity in just war theory, states bear duties to act in a way that causes the *least amount of harm* (relative to other protection efforts)—both for the human recipients of the operation and to the infrastructure of the adversaries.[41] This is true so long as it does not cause disproportionate harm to the intended beneficiaries or innocent bystanders.[42] Similarly, Article 57 of the First Additional Protocol of the Geneva Conventions states that when there is a choice between options, "the objective to be selected shall be that the attack on which may be expected to cause the least danger to civilian lives and to civilian objects."[43] The harm incurred by those targeted by CMOs is far less than in other more serious incursions and the cyber systems themselves remain unaffected (unlike with Stuxnet). Additionally, because of the nature of cyberspace and the Internet, cyber spies (unlike their HUMINT counterparts) can operate at a distance, away from imminent danger, and can anonymize themselves using virtual private networks—or at least have the option to do so. The risk, then, of being unmasked and incurring a serious threat to one's life or personhood is arguably less in the case of cyber spies than traditional HUMINT operators. The almost complete anonymity afforded to cyber spies is something that is not possible with HUMINT.

Let us assume that the cyber spy or cyber traitor (recall: a double agent employed both by the genocide regime and intervening state) is operating physically in-country. Should the cyber spy or cyber traitor become aware of the extermination camp blueprints and have the opportunity and resources to falsify them, an argument could be made that the cyber spy ought to relocate to a safe place to conduct the cyber operation. The spy or traitor could therefore travel to the state that is employing her for treasonous activities and conduct the CMO from afar. As Fabre argues, traitors and their families may be owed a duty of care to be "spirit[ed]" out of the dangerous territory should their safety be unjustifiably compromised as a result of their permissible treason.[44] If it is not feasible for the traitor to leave safely, or if conducting CMO is not possible from a distance because the system is air gapped (that is, not connected to the Internet), for example, then the concern remains. But there is, all things considered, at least a greater chance of cyber spies avoiding the risks attendant (albeit justly assumed) on manipulating an enemy's intelligence vis-à-vis their HUMINT counterparts. So, when considering how to achieve a particular end,

https://doi.org/10.1017/S0892679423000187 Published online by Cambridge University Press

states may have a duty to use CMOs over (or at least prior to) conventional HUMINT operations.

Finally, CMOs are likely to be more *cost effective*—monetarily—compared to conventional protection measures (such as military interventions or sanctions), which require a great amount of logistical coordination, human power, and resources. CMOs, which can feasibly be waged by one individual, are not likely to be as costly as such efforts.

The public (of the intervening state) need not know *what precise* CMOs have taken place to parry atrocities; after all, disclosures may compromise such operations and motivate the genocidal regime to patch relevant vulnerabilities in their networks. But, given the above factors (wherein CMOs are likely to be comparatively less harmful, more cost-effective, and more timely), CMOs may even bolster political will *within governments* for the prevention of atrocities. Therefore, CMOs may help mitigate much of the political paralysis that arises when states are confronted with evidence of atrocities. This is not to say, however, that using CMOs as part of CYBINT does not give rise to serious concerns.

Considerations

One potential objection to CMOs is that—by going further than merely stealing the information, as is done in strictly CYBINT—the operation increases the risk of the spy or mission being compromised. Through their cyber activities, the spy and the system's vulnerability may be detected. Detection is problematic on two fronts: (1) it severs cyber "back door" access, thereby precluding future access via that vulnerability, and (2) it increases the risk to the cyber spy and cyber traitor, whether based in-country or not. Let us begin with the latter concern.

First, cyber operations afford a large degree of anonymity for the person with her hands on the keyboard, thereby frustrating attribution efforts. Yet it is not beyond the realm of possibility that the genocidal regime might be able to trace the whereabouts or the identity of the cyber *spy*—even if she is based outside the target state (as her being in-state might not be required). The risk of being "found out" is higher if the CMO is conducted by a cyber *traitor*—someone internal to the regime who might be the only person who has the particular access necessary for the operation (or is more readily available to access the information undetected). Fabre, for instance, discusses the notion of "mandatory treason": the moral obligation for individuals to betray their political community by

https://doi.org/10.1017/S0892679423000187 Published online by Cambridge University Press

disclosing secrets that would help stymie fundamental rights violations by the enemy.[45] At the same time, she maintains that if there is a "high risk of [the traitor] being executed, tortured, or sentenced to a lengthy prison sentence"[46] should she be compromised, then the level of this risk would be unacceptable to the traitor. The traitor is under no obligation to disclose or—as I would argue—*manipulate* the content via cyberspace. Doing so may be permissible and a supererogatory act (all things considered). But it could be an unreasonable expectation to make of the cyber traitor.

Nonetheless, when discussing mandatory treason and traitors, Fabre contends that the extent to which the individual has a hand, causally or morally, in the rights violations she is supplying information about affects the cost she ought to accept.[47] Let us revise the "Blueprints for the Extermination Camp" scenario case slightly. Suppose it is not possible (or would be too time intensive) for a cyber spy in State *X* to hack into the network of the genocidal regime (State *Y*) and manipulate the data, and that, instead, State *X* has a cyber traitor who could do the operation in a more timely and effective manner. Say the cyber traitor herself helped design the camp prior to her "crossing over."[48] She would be obligated to conduct the CMO—and to shoulder the concomitant risks—more than a cyber spy who happened across the information. Following Fabre's logic, "An official who is partly responsible for rights violations is under a duty to act treasonably, whereas an ordinary citizen is not."[49] Moreover, if the cyber traitor enjoys a more privileged position, then perhaps she is more likely to be protected and get away with the cyber operation because she can cover up the activity (in our case, not mere CYBINT, but CMOs).[50]

As for the former objection, I suggest taking heed from the British government's response to solving the Enigma code. Upon cracking the code in 1941, the Allies refrained from acting on Nazi intelligence so as not to alert the regime that their code had been broken. This decision has been credited with having shortened World War II by two to four years.[51] Thus, if it is more valuable to continue conducting *strictly* CYBINT (because it may result in the genocide ending more quickly, for instance), then CMOs (which would foreseeably alert the regime of a breach) ought to be avoided. Indeed, it may be more permissible to act upon the intelligence gleaned from CYBINT and refrain from CMOs. For example, in the "Extrajudicial Killing" scenario:

> State *Z*'s CYBINT discovers State *Y*'s planned execution of prisoners of war and civilians, which will take place at Location *L* on Wednesday at 15:00. Based on this

https://doi.org/10.1017/S0892679423000187 Published online by Cambridge University Press

CYBINT, State *Z* plans to conduct a military intervention to prevent the war crime. It is crucial that the information resident in the atrocity perpetrators' networks is not altered, so that the executioners go to the correct location at the correct time (to then be ambushed by the interveners).

In this instance, the intelligence ought to remain unaltered. This is because the protective action relies on *not* manipulating the date, time, or location of the extrajudicial killing via CMOs. If, however, no other protection mission is forthcoming (due to unwillingness or inability), then this is precisely when CMOs should be considered. Altering or deleting the intelligence via CMOs in the "Extrajudicial Killing" scenario, for instance, might "buy time" for prisoners and civilians to escape, or to be liberated at the war's end.

Of course, CMOs may give rise to unintended consequences. The most extreme, perhaps, may be that the perpetrators expedite the genocide should such CMOs be discovered (conceivably out of rage for having been deceived and an acute awareness that their networks have been breached and so time to execute the atrocities may be perceived as limited). Similarly: having realized the orders to lay down arms were fabricated, perpetrators may come to believe that *all* orders, including those genuine calls from the head of state for a ceasefire, are fake.[52] For instance, is it permissible to deploy "deepfakes"—"media (including images, audio and video) that is either manipulated or *wholly generated* by [artificial intelligence]"—depicting genocidal leaders announcing that certain groups ought not be targeted (even if for just a short period of time)?[53] As an example, consider the potential release of a deepfake of Russian president Vladimir Putin declaring an armistice in the on-going invasion of Ukraine (where there is mounting evidence of war crimes).[54] Might this be permissible, or indeed even morally required?

On the one hand, due to the reach of the head of state, releasing false orders or a deepfake audio may result in a widespread pause in violence—not just the localized suspension of killing. It may also precipitate confusion among perpetrators as they scramble to ascertain the authenticity of their orders, resulting in some abandoning their posts, or captured victims being let free.

On the other hand, the falsification of a head of state's orders might muddy the powerful potential for diplomatic negotiations (as the target would—conceivably—not take kindly to being impersonated). Further (as noted above), if the head of state, as the "legitimate authority," *actually* calls for the full cessation of an atrocity, not all perpetrators may heed such orders, suspecting them to also

https://doi.org/10.1017/S0892679423000187 Published online by Cambridge University Press

be fabricated. (There are also personal harms that may be experienced by the head of state [like Putin] resulting from the deepfake; however, I contend genocidal leaders are liable to any such harms.)[55] On balance, then, I argue that the head of state ought to be exempt from impersonation via CMOs. It is paramount to preserve his legitimate authority, precisely because he (in this case, Putin) is the only person who can order an end to the atrocities. Conversely, middle- and lower-ranked military or government leaders may be justly impersonated; more specifically, the threshold for the impersonation of such ranks (as in the case of the false emails sent to Atroshchenko's wife, described above) is lower than for the head of state, precisely because their influence is not as far-reaching.

A further concern is that the manipulated content may be picked up *and believed to be authentic intelligence* by the intervening state's allies conducting their own espionage activities to prevent atrocities. The allies may in turn form plans based on (unbeknownst to them) incorrect information, or CMOs may contradict one another, undermining the whole enterprise. What is required, then, is "deconfliction": minimizing the potential for overlapping operations that jeopardize the mission.[56] To do so, I suggest developing, designing, or deploying CMOs in confidence with allies. Of course, there is a risk that sharing details of a CMO with allies increases the likelihood of a leak. Such is the risk of any information sharing between allies. If there is reason for suspicion, then perhaps the state ought to merely flag to its allies that certain systems, communications, or structures *may* be the target of CMOs (without sharing details).

In sum, I argue it is sometimes permissible, if not required, for spies to use CMOs to falsify a genocidal regime's own atrocity preparations or plans (as per the "Blueprints for the Extermination Camp" scenario). Misleading, tricking, or confusing potential perpetrators by sending fictitious orders (as per the "Email" scenario) and planting false (or contradictory) information is also pro tanto permissible. In this case, CMOs must: (1) seem likely to prevent violence at no supererogatory cost to one's own spies or regime, (2) be discriminate, (3) be less harmful than traditional protection measures (such as severe sanctions or armed interventions), and (4) not imitate the head of state, thereby keeping open the avenue for a true cessation of atrocities to be ordered. Crucially, and at the very least, CMOs must (5) foreseeably "buy time" for potential interveners to prepare other protection efforts or for perpetrators to "think twice" about their atrocity plans.

https://doi.org/10.1017/S0892679423000187 Published online by Cambridge University Press

Conclusion

In this essay, I have focused on the ethics of using cyber manipulation as part of CYBINT. I considered using cyber operations to manipulate the enemy's own intelligence—pertinent to committing atrocities—resident in their cyber systems; that is, to (in Tzu's words) "carry false tidings to the enemy"[57] *about the enemy* to prevent atrocities. I concluded that CMOs as part of CYBINT are not only pro tanto permissible but also occasionally required, especially over (or at least prior to) a HUMINT operation that aims to achieve the same end.

This essay has only scratched the surface of a deep topic, and many questions regarding the ethics of CMOs remain. For example, who, exactly (other than states), may permissibly launch CMOs to prevent atrocities? May nongovernmental organizations, technology corporations, or hacktivist groups like Anonymous conduct CMOs?[58] In this essay, I have focused on cyber spies and cyber traitors acting with explicit orders from a state's intelligence agency; but what if, due to time constraints, severed or insecure communication lines, or concerns regarding "moles," no order for cyber manipulation can be made? Should individuals independently wage their own CMOs, as a form of permissible vigilantism? Further still, might it be permissible, or even required, to launch online *disinformation* campaigns across social media (Facebook, Twitter, TikTok, and so on) to deter perpetrators from committing atrocities?[59] These questions are avenues for future research. The aim of this essay has been to explore the ethics of cyber manipulation operations and to suggest that states would do well to employ a pseudo "Agent Zigzag 2.0" to parry atrocities—that is, of course, if they do not already.

Notes

[1] Michael Hayden, quoted in Ben Buchanan, *The Cybersecurity Dilemma: Hacking, Trust, and Fear between Nations* (London: Hurst, 2016), p. 287, fn. 14.

[2] For more on Agent Zigzag, see Ben Macintyre, *Agent Zigzag: The True Wartime Story of Eddie Chapman; Lover, Traitor, Hero, Spy* (London: Bloomsbury, 2007).

[3] See Cécile Fabre, *Spying through a Glass Darkly: The Ethics of Espionage and Counter-Intelligence* (Oxford: Oxford University Press, 2022), p. 99.

[4] Sun Tzu, *The Art of War*, trans. Lionel Giles (sixth century BC; repr. London: Luzac and Co., 1910), §23, p. 30.

[5] On "doomed spies," Sun Tzu suggests "doing certain things openly for purposes of deception, and allowing our spies to know of them and report them to the enemy." Ibid., §12, p. 29.

[6] Elias Groll, "Cyber Spying Is Out, Cyber Lying Is In," *Foreign Policy*, November 20, 2015, foreignpolicy.com/2015/11/20/u-s-fears-hackers-will-manipulate-data-not-just-steal-it.

[7] Other scenarios of unjust policies or violations of foundational rights might plausibly also qualify; however, these fall beyond the scope of this investigation. I am also not concerned with so-called canary traps—the tactic of releasing false information to see which is leaked and thereby identifying the leak. See Tom Clancy, *Patriot Games* (New York: Putnam, 1987).

https://doi.org/10.1017/S0892679423000187 Published online by Cambridge University Press

[8] It is not that such efforts must be exhausted, but that there must be an attempt to dissuade individuals via honest means first.
[9] Sissela Bok, *Lying: Moral Choice in Public and Private Life* (New York: Random House, 1978), p. 31.
[10] Fabre, *Spying through a Glass Darkly*, pp. 16, 86.
[11] Ibid, p. 86.
[12] Ibid., p. 155.
[13] Fabre, *Spying through a Glass Darkly*, p. 7.
[14] Ibid., p. 174.
[15] Fabre, *Spying through a Glass Darkly*, p. 227. References to proportionality, discrimination, and necessity echo that of the just war principles that govern war.
[16] For more on the ethics of intelligence, see Toni Erskine, "'As Rays of Light to the Human Soul'? Moral Agents and Intelligence Gathering," *Intelligence & National Security* 19, no. 2 (2004), pp. 359–81, at pp. 363–65; Ross Bellaby, "What's the Harm? The Ethics of Intelligence Collection," *Intelligence and National Security* 27, no. 1 (2012), pp. 93–117, at p. 93; and Ross W. Bellaby, "Justifying Cyber-Intelligence?," *Journal of Military Ethics* 15, no. 4 (2016), pp. 299–319, at p. 299.
[17] United Nations General Assembly, "2005 World Summit Outcome," A/RES/60/01, September 16, 2005, paras. 138–39.
[18] On the ethics of hacking into nonliable networks, see Fabre, *Spying through a Glass Darkly*, p. 183.
[19] Ibid., p. 182.
[20] I am not claiming that covert or espionage operations are less democratically accountable; I am simply suggesting that it would not be prudent for such cyber operations to be openly deliberated *prior to* their operationalization, as that would foreseeably undermine their effectiveness. The CMO could be disclosed to the democratic public after the fact, so long as it does not undermine the efficacy of the operation or the safety of its operatives.
[21] Fabre, *Spying through a Glass Darkly*, pp. 191–92.
[22] Joint Chiefs of Staff, *DOD Dictionary of Military and Associated Terms* (Washington, D.C.: Department of Defense, June 2020), p. 55.
[23] I am grateful to Juan Espindola for raising this point.
[24] For instance, see "immunity asset" and "sexual blackmail" in Fabre, *Spying through a Glass Darkly*, p. 145.
[25] Ibid., pp. 185–87.
[26] Barbara Starr, "U.S. E-Mail Attack Targets Key Iraqis," CNN.com, January 12, 2003, edition.cnn.com/2003/WORLD/meast/01/11/sproject.irq.email/index.html; and Richard A. Clarke and Robert K. Knake, *Cyber War: The Next Threat to National Security and What to Do about It* (New York: Ecco, 2012), p. 10. See also Shaheed Nick Mohammed, *The (Dis)information Age: The Persistence of Ignorance* (New York: Peter Lang, 2012), p. 88.
[27] Starr, "U.S. E-Mail Attack Targets Key Iraqis."
[28] Ibid.
[29] Clarke and Knake, *Cyber War*, p. 10.
[30] Ibid.
[31] InformNapalm, "Hacking of a Russian war criminal, commander of Military Unit 75387, 960th Assault Aviation Regiment," *InformNapalm,* March 27, 2023, informnapalm.org/ua/zlam-75387-960-aviapolku/.
[32] Rhiannon Neilsen, "'Honey, I'm Hacked': Ethical Questions Raised by Ukrainian Cyber Deception of Russian Military Wives," *Just Security,* May 18, 2023, www.justsecurity.org/86548/honey-im-hacked-ethical-questions-raised-by-ukrainian-cyber-deception-of-russian-military-wives/.
[33] There are distinct ethical considerations attendant with this example that do not apply here; see ibid.
[34] Steven Whitham, Army Cyber Institute, West Point, interviewed by Rhiannon Neilsen, 2018.
[35] Ibid.
[36] Neil C. Rowe, "The Ethics of Cyberweapons in Warfare," *International Journal of Cyberethics* 1, no. 1 (2009), p. 12.
[37] Ibid., p. 12. See also Jon R. Lindsay, "Stuxnet and the Limits of Cyber Warfare," *Security Studies* 22, no. 3 (2013), pp. 365–404, at p. 384.
[38] On using cyber operations to create a feedback loop, see Clarke and Knake, *Cyber War*, p. 7; and Lindsay, "Stuxnet and the Limits of Cyber Warfare," p. 384.
[39] Toni Erskine, "Existential Threats, Shared Responsibility, and Australia's Role in 'Coalitions of the Obligated,'" *Australian Journal of International Affairs* 76, no. 2 (2022), pp. 130–37; and Toni Erskine, "Moral Agents of Protection and Supplementary Responsibilities to Protect," in Alex Bellamy and Tim Dunne (eds.), *The Oxford Handbook of the Responsibility to Protect* (Oxford: Oxford University Press, 2016), pp. 167–85, at p. 180.

[40] On the ethics of hacking into cyber networks more broadly, including cyber operations that target non-liable individuals and networks, see Fabre, *Spying through a Glass Darkly*, p. 183.
[41] See Seth Lazar, "War," in *The Stanford Encyclopedia of Philosophy* archive, ed. Edward N. Zalta, Spring 2020 edition, plato.stanford.edu/archives/spr2020/entries/war/.
[42] Cécile Fabre, *Cosmopolitan War* (Oxford: Oxford University Press, 2012), p. 203; and Daniel Baer, "The Ultimate Sacrifice and the Ethics of Humanitarian Intervention," *Review of International Studies* 37, no. 1 (January 2011), pp. 301–26, at p. 315.
[43] Art. 57(3), "Precautions in Attack," in Diplomatic Conference on the Reaffirmation and Development of International Humanitarian Law Applicable in Armed Conflicts, "Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of Non-International Armed Conflicts (Protocol I)," June 8, 1977.
[44] Fabre, *Spying through a Glass Darkly,* p. 127.
[45] Ibid., pp. 123–26.
[46] Ibid., p. 132.
[47] Ibid.
[48] Think Karl Bischoff and Fritz Ertl, who designed Auschwitz-Birkenau. See Harold Marcuse, "Architecture and Auschwitz," *Journal of Architectural Education* 49, no. 2 (November 1995), pp. 123–28, at p. 125.
[49] Fabre, *Spying through a Glass Darkly*, p. 133, fn. 33.
[50] Ibid.
[51] Jack Copeland, "Alan Turing: The Codebreaker Who Saved 'Millions of Lives,'" BBC News, June 19, 2012, www.bbc.com/news/technology-18419691.
[52] I am grateful to Cécile Fabre for raising this objection.
[53] Nina Schick, *Deepfakes: The Coming Infocalypse* (New York: Hachette, 2020), p. 1.
[54] United Nations General Assembly, "Independent International Commission of Inquiry on Ukraine," A/77/533, October 18, 2022; and Lorenzo Tondo, "Russia Has Committed War Crimes in Ukraine, Say UN Investigators," *Guardian*, September 23, 2022, www.theguardian.com/world/2022/sep/23/russia-has-committed-war-crimes-in-ukraine-say-un-investigators.
[55] For more on personal harms, see Regina Rini and Leah Cohen, "Deepfakes, Deep Harms," *Journal of Ethics & Social Philosophy* 22, no. 2 (July 26, 2022), pp. 143–61.
[56] My thanks go to Herb Lin for bringing this term to my attention.
[57] Tzu, *The Art of War,* p. 30.
[58] James Purtill, "Hacker Collective Anonymous Declares 'Cyber War' against Russia, Disables State News Website" ABC News, updated February 25, 2022, www.abc.net.au/news/science/2022-02-25/hacker-collective-anonymous-declares-cyber-war-against-russia/100861160.
[59] "The Responsibility to Deceive? Deploying Online Disinformation & 'Fake News' for Atrocity Prevention," YouTube video, 59:13, a talk given by Rhiannon Neilsen, posted by CISAC (Center for International Security and Cooperation), Stanford University, March 7, 2023, www.youtube.com/watch?v=7z3xm7v3sJo&t=1s.

Abstract: Intelligence operations overwhelmingly focus on obtaining secrets (espionage) and the unauthorized disclosure of secrets by a public official in one political community to another (treason). It is generally understood that the principal responsibility of spies is to successfully procure secrets about the enemy. Yet, in this essay, I ask: Are spies and traitors ethically justified in using cyber operations not merely to acquire secrets (cyber espionage) but also to covertly manipulate or falsify information (cyber manipulation) to prevent atrocities? I suggest that using cyber manipulation operations to parry atrocities is pro tanto morally permissible and, on occasion, required.

Keywords: cyber, espionage, ethics, genocide, atrocities, information operations, influence, intelligence