The Moral Permissibility of Automated Responses during Cyberwarfare

David Danks
Department of Philosophy, Carnegie Mellon University; and
Institute for Human & Machine Cognition

Joseph H. Danks
Center for Advanced Study of Language, University of Maryland, College Park

Abstract

Automated responses are an inevitable aspect of cyberwarfare, but there has not been a systematic treatment of the conditions in which they are morally permissible. We argue that there are three substantial barriers to the moral permissibility of an automated response: the attribution, chain reaction, and projection bias problems. Moreover, these three challenges together provide a set of operational tests that can be used to assess the moral permissibility of a particular automated response in a specific situation. Defensive automated responses will almost always pass all three challenges, while offensive automated responses typically face a substantial positive burden in order to overcome the chain reaction and projection bias challenges. Perhaps the most interesting cases arise in the middle ground between cyberoffense and cyberdefense, such as automated cyberexploitation responses. In those situations, much depends on the finer details of the response, the context, and the adversary. Importantly, however, the operationalizations of the three challenges provide a clear guide for decision-makers to assess the moral permissibility of automated responses that could potentially be implemented.

Keywords

Automated responses; cyberwarfare; cyberdefense; cyberexploitation; attribution problem; projection bias; chain reactions; moral permissibility

1. Introduction

One striking difference between warfare in the cyber and kinetic domains is the variation in timescales on which they proceed, and so the different types of responses that are suitable, or even possible. Events in cyberwarfare—attacks, defenses, counter-attacks—can occur on the scale of seconds or even faster, whereas most events in kinetic warfare may unfold more slowly over days, weeks, or even months. Since humans cannot respond sufficiently quickly to all relevant cyberwar events, successful cyberwarfare requires the (non-exclusive) use of automated responses: our cyber network systems must include at least some policies of the form "If event *E* happens in context *C*, then execute response *R*" (where *E*, *C*, and *R* can all be highly complex), where *R* is performed whenever the triggers—combinations of specific *E*s in particular *C*s—occur without the need, or perhaps even the possibility, of any human intervention. For example, if an intrusion to a computer network is detected with a specific signature (possibly attributable to a specific adversary), then a system might automatically respond in various ways (possibly tailored to that specific intrusion/adversary).

In this paper, we examine the ethical permissibility of various types of automated responses. Debates about the morality of acts in warfare have principally focused on actions or responses that proceed immediately from deliberation by relevant human decision-makers. In contrast, we are principally concerned with the conditions for the moral legitimacy of automated responses that occur without any human decision at the moment of initiation.[1] In particular, we present three different barriers—each of which can be made relatively operationally clear—that automated responses must overcome in order to be ethically permissible. Previous arguments

[1] And because we are interested in automated actions, we focus solely on responses. It is unclear whether it even makes sense to talk about an "automated" unprovoked action.

about automated responses have focused on only one of these (the *attribution challenge*). We argue that the ethical permissibility of automated responses depends also on the (in)ability to predict rapid sequences of events that can result from the use of automated responses (the *chain reaction challenge*), and on human decision-makers' (in)ability to accurately predict their future preferences and desires (the *projection bias challenge*). These challenges or tests can be used to draw important distinctions between classes of automated responses, as well as help to understand the ethical permissibility of activities in computer network operations (CNO)—in particular, computer network exploitation (CNE or cyberexploitation)—that occupy an ambiguous middle ground between computer network defense (CND or cyberdefense) and computer network attacks (CNA or cyberoffense).

Any type of warfare requires conditional planning in which decision-makers determine which actions or responses to take if certain events happen in a particular context. In many cases, the context *C*, triggering events *E*, and response *R* can be quite complex. Mission plans in kinetic warfare, for example, often include long lists of contingency plans for the many different possible scenarios that might be encountered. Many of these conditionally triggered actions are publically declared in advance, as in mutually assured destruction plans during the Cold War or mutual assistance clauses in treaties. Although all warfare involves conditional plans, cyberwarfare is distinctive because at least some of those conditional plans will need to be automated in order to function properly on the compressed timescales of cyberevents and are unlikely to publically declared in advance. That is, some of these conditional plans will need to be implemented so that *R* is automatically triggered whenever the antecedent is satisfied without the need or possibility of human intervention or decision, and without prior notification of the adversary. Few analogues to such automated responses occur in the kinetic domain. Mission plans and wargaming produce conditional plans, but those are almost always implemented by human decision-makers who can

change them "on the fly" when *C* changes, which is especially important when *C* includes many factors. The closest analogues are doomsday triggers for nuclear weapons that guarantee a nuclear strike whenever certain preconditions obtain, regardless of whether a human is present to launch the (counter-)attack. These are completely different from automated responses in the cyber domain, however, as doomsday triggers are publicly declared and involve responses *R* constituted by overwhelming (kinetic) force. In contrast, automated cyber-responses are rarely publicly declared, and can involve a much broader range of responses. They thus warrant a closer investigation, particularly into their ethical permissibility.

We focus in this paper on the response *R*, though we recognize that there are many interesting questions associated with *C* and *E* (e.g., does the language used to describe *C* and *E* create operational "blind spots" by grouping together contexts or events that are actually different?).[2] In Section 2, we provide three key challenges for the ethical permissibility of any automated response. These challenges can be operationalized into tests that help to reveal ethical differences within the broad class of automated responses. In general, we distinguish between offensive and defensive responses based on the extent to which the response aims to destroy, damage, or otherwise adversely impact another party's resources and capabilities, not just in computer network operations (CNO), but in their infrastructure and kinetic capacities as well. For example, launching a Denial of Service (DoS) attack is an offensive response, while simply closing a server port is a defensive one. There is obviously not a "bright line" distinction here, as responses can be more or less offensive or defensive. For example, cyberexploitation is defensive according to this definition, but in a larger sense CNE can clearly be an important part of a

[2] There are also interesting questions about the combatant status of various individuals involved in the cyberwarfare process. For example, how should we think about someone who writes code but does not actually put it into action? Is such an individual comparable to a worker in a munitions factory, or a military support individual, or some other role? For space reasons, we do not address those issues in this paper.

subsequent offensive response (e.g., providing reconnaissance information for a cyberattack, or installing a "back door" to the adversary's systems that is harmful only if triggered in the future). This distinction is nonetheless a useful one that is reflected in both language (the very terms 'cyberattack' vs. 'cyberdefense') and allocation of equities and authorities between various federal agencies. We apply the three operational tests (outlined in the next section) in Sections 3 and 4 to various cases drawn from this cyberoffense-to-cyberdefense spectrum, including cyberexploitation, to show how their moral permissibility can depend on details of the cyber-action, the adversary, and the broader context.

2. Three challenges

We focus throughout on the ethical permissibility of automated responses that must occur without the possibility of human intervention.[3] Obviously, some automated responses will be slow enough that that they could be overridden by human decision-makers, but we are principally interested in those that must be established prior to being in the actual triggering conditions. The necessity of enacting them beforehand creates three natural challenges. The most commonly discussed one is the so-called *attribution challenge* or *problem* (e.g., Dipert, 2006, 2010). A necessary condition to be morally justified in response *R* against another party *P* is that one knows that *P* attempted or intended to harm in some way. More properly, one must have a highly justified belief that *P* is responsible for some deliberate attempted/intended harm *H* (including both violations of sovereignty and actual damage to one's resources) where the amount of justification

[3] We assume throughout that only humans, either individuals or groups of individuals, can (in these settings) be a locus of moral responsibility. In particular, we assume that computers cannot be morally responsible for their actions. We also take as given that these automated responses are at least *potentially* morally justified in particular situations, as the mere passage of time between the decision and the action is not sufficient (on its own) to eliminate the possibility of moral justification. We instead focus on when the potential justification is actual.

that is ethically required depends in part on the nature of both *R* and *H* (as well as other factors).[4] Of course, other conditions must also be satisfied for one to be morally justified in performing *R*, such as the requirements that *R* be proportional to *H*, that *R* be appropriately targeted towards only *P*, and so forth; Just War doctrine provides one way of spelling out these requirements (e.g., Walzer, 1977). But we focus here on the necessary attribution condition stated above, and so can set aside discussions about, for example, the best statement of Just War principles.

The oft-noted attribution challenge in the cyber domain is simply that attribution to *P* can be remarkably difficult. There are many cyber-tools and techniques for hiding one's identity, changing the apparent source of an attack, distributing an attack across multiple systems or agents, or otherwise disguising the party responsible for some (attempted or actual) *H*. Thus, one natural conclusion is that automated responses are rarely an option from an ethical point-of-view, since the attribution challenge precludes the possibility of ever having a sufficiently justified belief that *P* is responsible for *H* (see also Dipert, 2006, 2010 on the many complexities prompted by the attribution problem). Perhaps responses that have no possibility of harming the other party could avoid the attribution challenge, but it seems likely that this is a relatively small set, and determining the possibilities of future harm may be difficult, if not impossible.[5]

The attribution challenge is a very real issue in the cyber-domain, and has received the most attention in discussions of the ethical permissibility of automated responses (and even non-

---

[4] One might hope to have a function that takes *H* and *R* (and undoubtedly other factors) as input and then outputs an appropriate threshold of justification, perhaps expressed as a probability. This possibility seems unlikely for at least two reasons. First, the 'other factors' mentioned in the parenthetical are likely to be a hopelessly complex set, as almost anything can be relevant under sufficiently unusual circumstances. Second, any proposed threshold would likely be vulnerable to a *sorites*-type objection: if τ is acceptable, then $\tau - \epsilon$ will almost certainly be acceptable as well, and so we have a wedge to show that no threshold is defensible. More about the issue of what level of justification is required can be found in, e.g., Zimmerman (1997), Rosen (2003), and Guerrero (2007).

[5] Though see our discussion of defensive automated responses in the next section.

automated ones). Importantly, however, it is arguably significantly blunted when we focus on automated responses in cyber*warfare*, in contrast with simple cyber*attacks*. Cyberwarfare involves groups with the expertise and resources to mount a significant attack, including the accompanying research and development costs, and so arguably includes only those with the backing of a nation-state, whether the group is officially part of the state (e.g. military), or only sponsored (e.g., contractors), encouraged (e.g., patriotic hackers), or tolerated (e.g., international crime) by the state. State-backed groups engaged in cyberwarfare focus on more substantial intrusion sets or cyberattacks (e.g., Stuxnet or Flame; see Sanger, 2010) which require months, if not years, of research, development, and testing prior to deployment, and typically have a goal that serves the interests of a particular state or state-like group. There is a limited set of potential adversaries or threats, as well as a limited and informative set of attacker intents, such as significant physical harm, financial damage, disruption of infrastructure, or political interference. For example, Stuxnet apparently was intended to result in kinetic damage to Iranian uranium centrifuges, while the objective of Flame was exfiltration of intelligence information (Sanger, 2010).

The same research and development time and activity that enable the attacks to be much more sophisticated also provide the opportunity for discovery by cyber-defenders through cyberexploitation (or non-cyber methods). In particular, this development time can provide defenders with the opportunity to develop and implement attack recognition algorithms that can provide the required justification for an automated response that is presumably tailored to the particular attacker. This activity is analogous to military reconnaissance in preparation of the kinetic battle space, as both activities are intended to determine various possibilities so that one can readily and appropriately respond in the "fog of war" or, in the cyber-case, the "compressed timescale of war." Thus, the very development time required for cyberwarfare attacks also

provides the window for defenders to discover (whether through cyber- or other means) information about the possible attackers and intrusion signatures of possible attacks so that attribution is, while not guaranteed, significantly improved. As a result of all of these distinctive features of cyberwarfare, the attribution problem will, for good reasons, arise much less frequently in this context. It thus typically presents less of an impediment to morally justifiable automated responses, though it is still a challenge that must be met in particular situations.

There are two other, less-discussed challenges to the moral permissibility of automated responses. Moreover, both are arguably exacerbated in the cyberwarfare domain, though they present potential challenges to actions in the kinetic realm as well. The *chain reaction challenge* arises from the difficulty of knowing the likely direct and indirect effects of the response R. As noted above, R must be appropriate for the attempted/intended harm H, but it can be quite difficult in the cyber-domain to know whether R really is suitable, as R can trigger unexpected automated responses from the adversary or could interact in negative ways with other automated responses of the defenders. As a result, a feedback loop resulting in undesirable outcomes can rapidly result. For example, the attackers might have their own automated response of the form "If R, then do E_1," where E_1 represents a slight escalation of the situation. Suppose then that the defenders have an automated response "if E_1, then R_1" that further (slightly) escalates the situation. It is straightforward to see that a rapid-fire escalation could easily occur—the adversary doing E_1, …, E_n while we do R_1, …, R_n—such that the final actions are completely unjustified by the original action, but are also uncontrollable and unstoppable. The whole back-and-forth exchange could spiral out of control before any person could intervene.

One might object that this is a fanciful and implausible scenario, but a similar phenomenon has already been observed in financial transactions, such as when interactions between automated stock trading systems caused a rapid, unexpected, and uncontrolled drop in the Dow

Jones on May 6, 2010 (in which, for example, Proctor & Gamble stock dropped by over one-third in less than a minute, entirely because of a sequence of rapid interactions between automated stock trading systems). More recently, the introduction of a novel trading algorithm, apparently without a "stop" option, led to uncontrolled trades by Knight Capital Group, which in turn impacted the New York Stock Exchange and the Dow Jones industrial average (Popper, 2012). This problem can even arise in the kinetic warfare realm, though one hopes that there will be typically be time and opportunity to slow or stop such escalation when human decision-makers are involved.

The difficulty of knowing the indirect effects of *R* is proportional to the resources of the adversary, and so is particularly acute in the cyberwarfare context. Groups with significant cyber-resources are more likely to have sophisticated automated response systems of their own, and so the possibility of a chain reaction increases because of both the complexity and sophistication of the adversary's responses and also the simple increase in combinatorial possibilities. In addition, more resourceful adversaries are both able and more likely to change their automated response systems immediately before beginning an attack, precisely to make counter-attacks less likely to succeed. As a result, cyber-defenders are much less likely to have the knowledge required to predict the likely effects of their own automated responses. And since cyberwarfare involves a highly resourceful and determined adversary, the overall chain reaction challenge is particularly acute: the adversaries are highly likely to have systems in place about which the defenders know relatively little, but that could interact negatively with their automated responses. Of course, this is not an insurmountable challenge: given sufficient knowledge (from diverse sources), one could potentially determine the likely effects of *R* (though see Rowe, 2010). But it is a major barrier to the ethical permissibility of automated responses, particularly given the relative instability of cyberwarfare tactics.

The final challenge centers on the humans who decide to implement an automated response, and so are (or should be) ultimately morally responsible for it. The core problem is that people are not nearly as good as they think at making accurate predictions about their beliefs, desires, and preferences in novel contexts, especially future situations. The issue is that people suffer from *future self projection bias*[6]: they tend to believe that they will, in the future, be essentially the same as they are now, even though people change significantly as their environments and situations shift.[7] A large body of research in many different domains—everything from mountain climbing to lotteries to meals to medical decisions (see, e.g., the studies reviewed in Loewenstein, O'Donoghue, & Rabin, 2003)—has shown empirically that people's predictions about how they would respond in context *C* differ significantly from how they actually do respond when *C* comes about, where the predictions are worst when *C* is significantly novel, as *C* shares many fewer features with the current context (or familiar ones that they can readily imagine).

The *future self projection bias challenge* —for simplicity, just *projection bias challenge*—for automated responses is that the decision to implement an automated response must be made on the basis of predictions of exactly this sort. When implementing *any* conditional policy, one must think about how one will want to respond in context *C*, which requires predicting one's future preferences,

---

[6] In general, a projection bias is the (unjustified) attribution of one's own psychological attributes (beliefs, desires, etc.) to others. Projection bias often involves "mirroring" one's own perspectives onto an adversary, rather than adopting the adversaries' point of view. In this research literature, one's own future states are regarded as "other people" relative to one's current state, and so "future self projection bias" is a specific case of this more general bias.

[7] One particularly striking example that shows the pervasive nature of this phenomenon is that people who buy winter jackets on cold days are more likely to subsequently return those jackets (Conlin, O'Donoghue, & Vogelsang, 2007). Why? Because they are cold at the moment of purchase and so expect to be cold in the future, and thus overestimate the likelihood that they will use the jacket in the future. (And items that people do not use as frequently as they expect are more likely to be returned.) A second example comes from the stability of older adults' preferences for life-sustaining medical treatment (Ditto et al., 2003). Such preferences were collected from elderly adults three times at yearly intervals. Their preferences on whether to receive a treatment or not were only moderately stable—about 70% were the same each year as the year before. So people cannot anticipate their preferences even for serious life decisions.

beliefs, and desires. Cyberwarfare is arguably a highly novel context, and so all of this research implies that people have significant reason to doubt their ability to accurately predict the requisite preferences and intentions.

This inability is morally problematic because defenders could have an automated response that was deliberately selected and implemented, but which, at the moment of its execution, no decision-maker actually wants to have occur; they *thought* that they would want that response, but like so many predictions about one's own future states, they were wrong. Thus it's possible to have an act of cyberwarfare by a group that is deliberate (i.e., not accidental), but which no member of the group endorses or desires.[8] A standard necessary (though of course not sufficient) condition for moral responsibility for an action *A* is that the agent endorses or desires *A* (e.g., Frankfurt, 1969; Pereboom, 2000; Wolf, 1987); one is not morally responsible for unintended accidents (assuming they were also unforeseeable). But automated responses can thus easily be deliberate acts of cyberwarfare for which no one is responsible. As a result, there should be a (defeasible) moral presumption against such automated responses, as they have the clear potential to be acts that violate long-standing aspects of the ethics of warfare, principally that there should be decision-makers who are morally responsible for each deliberate act of warfare. To defeat the projection bias challenge, one must show that a particular decision to implement an automated response does not present a substantial risk of being such a violation.

There are two potential, generally mitigating factors for this particular challenge, though both are dependent on empirical data that (to our knowledge) have not yet been collected. First, many decisions in the cyberwarfare context are made by groups, whether small teams or large policy-making bodies. In contrast, the psychological literature examining people's (in)ability to

[8] Of course, a similar possibility exists in the kinetic realm. However, those cases are much more likely to involve the possibility or requirement of action at the moment, and so those actions would presumably be better aligned with actual decision-maker preferences.

predict their future preferences and desires has focused exclusively on individuals, and so little is known about whether groups would exhibit these biases. Psychological research on team and group decision-making has shown that groups can sometimes find better solutions than individuals, but can also be trapped into suboptimal outcomes (e.g., Campbell, 1968; Hirokawa, 1980; Woolley, Gerbasi, Chabris, Kosslyn, & Hackman, 2008). Much depends on the particular internal dynamics of the group, and little is known about the impact of those dynamics on possible future self project biases. That is, we (as a scientific community) simply do not know whether groups deciding to implement automated responses in a cyberwarfare context are subject to future self projection bias.

The second potentially mitigating factor is that many decisions to implement particular automated responses arise from red team-blue team activities in which the relevant decision-makers run through a range of possible scenarios to determine which responses are likely to be successful and which are not. These exercises also help to identify weak points in one's cyber-security that could perhaps be addressed with automated responses. To the extent that red team-blue team activities succeed in placing participants into cognitive states that are relevantly similar to the actual contexts, the decisions made during them should not be subject to the projection bias challenge. The problem is that the relevant realism of these exercises is simply unknown, since it must be realistic in the minds of the participants, rather than in technology, displays, and so forth. That is, the people participating in red team-blue team exercises must be sufficiently engaged in the exercise that they actually manage to place themselves (at least, temporarily) into the relevant context. To the extent that people recognize that it is "just an exercise" and so maintain cognitive distance, it is unlikely that they will avoid the biases that underlie the challenge. In fact, many studies have asked people to "imagine" or "visualize" themselves in the novel context *C* before giving their predictions, and those studies have still found that people

have great difficulty predicting their future preferences and desires (e.g., Read & Van Leeuwen, 1998). Thus an open empirical question is whether red team-blue team activities are sufficient to overcome the projection bias challenge.

3. The "easy" extremes

The previous section laid out three general challenges to the ethical permissibility of automated responses. We now turn to the problem of how to judge these challenges in some classes of cases, beginning with the extremes of the offense-to-defense spectrum. We take it to be uncontroversial that completely defensive automated responses, such as blocking an attacking intrusion or diverting it to a harmless site, are almost always ethically justified in terms of their impacts on others, including adversaries. In general, actions that *directly influence* only cyber-resources that are under our *legitimate control* will always be permissible in terms of the warfare context (though not necessarily well-advised), regardless of whether they are triggered automatically or implemented by a human. Moreover, this intuition accords with the three challenges outlined in the previous section. The attribution challenge is irrelevant: the response does not directly affect any other systems, so defenders do not (in the extreme case) need to worry about whether their attack attribution was correct.

The chain reaction and projection bias challenges are slightly trickier, but not significantly so. It is possible that an automated response that directly influences the defenders' own computer network systems would trigger a chain reaction involving the adversary (e.g., if the adversary's responses depended on internal features of the defenders' computer network systems), but it is hard to construct a scenario in which the defenders would be responsible for such a chain reaction. The fault would lie with the adversary because (by assumption) the adversary does not have legitimate claim to use (or depend on) the defenders' computer network systems. And

although a chain reaction involving only the defenders' own systems might occur and be disastrous in its effects, it would not be ethically problematic in terms of the warfare context. Of course, such a chain reaction could be morally problematic for other reasons (e.g., if it shut down the IT systems for the defenders' hospitals), but we leave those considerations aside. Similarly, it is possible—perhaps even likely—that projection bias will be an issue for our decision-making, and that the defenders might well regret the automated response. But that regret would not be over an act of *warfare* for which no one was responsible, assuming the automated response only directly influences systems under the defenders' legitimate control.

Of course, it can sometimes be unclear whether a particular automated cyber-response falls into this class, as both 'directly influence' and 'legitimate control' are notions with fuzzy boundaries. With regards to direct influence, sending back corrupted information to the adversary might appear to not directly influence the adversary, except that the manner of corruption could (unknowingly) lead to adverse events in the adversary's systems. Alternately, if system *S* is necessary for system *T*'s proper functioning, then actions that directly influence only *S* occupy a vague area between direct and indirect influence with regards to system *T*. Along the dimension of legitimate control, the interdependence of networks can easily lead to situations in which it is unclear whether one can legitimately exert influence on a cyber-system. For example, if there is an official agreement linking systems *A* and *B* with distinct owners, then the legitimate owners of *A* can justifiably claim some authority over *B* by virtue of the legal linkage between the two systems.

An example can show the complexities that can arise when trying to decide whether an automated response is purely defensive. One might think that defenders are always permitted to respond to a presumptive DoS or DDoS attack by ignoring requests from a particular range of IP addresses, as ignoring requests seems to directly influence only the defenders' own machines (as

they are simply telling them not to acknowledge certain packets). Suppose, however, that this particular server provides (previously negotiated) crucial support for a hospital's critical infrastructure. In that case, the defenders are not necessarily justified in ignoring messages from machines in that hospital, even if it appears that some of those machines are beginning a DoS attack on computer systems for which the defenders are responsible. Instead, they face the additional burden of determining whether the users of those machines are knowing or unwitting participants in the DoS attack. The issue becomes further complicated if a botnet is used to implement a DDoS, where some-but-not-all of the machines in the botnet are part of the critical infrastructure for an organization. Nonetheless, despite these complications, defensive automated responses—at least, the obvious cases—are less problematic from an ethical point-of-view as they focus on internal cyber-resources over which the defenders have legitimate control.

Offensive automated responses are more ethically complicated. We follow Owens, Dam, and Lin (2009) in understanding offensive cyber-actions (i.e., cyberattacks) to be "deliberate actions to alter, disrupt, deceive, degrade, or destroy computer systems or networks or the information and/or programs resident in or transiting these systems or networks" (p. 9). A key element here is that the attacker does not have legitimate control over the target systems. We further restrict our attention to offensive cyber-*responses*—cyberattacks prompted by outside actions or events, and typically intended to be retaliatory in nature. The attribution challenge is obviously relevant for such automated responses, as defenders may ethically inflict harm only on those that justly deserve it. However, as we argued in the previous section, this challenge is unlikely to be a substantial barrier in the cyberwarfare context, precisely because there are many more opportunities to learn features of the adversary's attack (e.g., the attack signature) that enable rapid, even automatic, attribution. The chain reaction and projection bias challenges are less easily handled.

In cyberwarfare, adversaries will almost certainly have the resources to have measures designed to protect their own computer networks. If their measures are all purely defensive, then the chain reaction challenge is unlikely to be a substantial barrier, as actions "internal" to the adversary are unlikely to trigger a further response from the defenders. Of course, the same resources that enable adversaries to protect their systems also make it substantially more likely that some of those automated counter-responses will be offensive in nature. The defenders may even have positive knowledge (whether through cyber- or other channels) that the adversary does have offensive automated counter-responses in place. That knowledge will, however, rarely translate into knowledge of the precise adversary counter-responses, and so defenders will rarely know anything more than "a chain reaction is highly possible."

In general, people have a positive ethical obligation to determine the likely direct and indirect effects of their actions, and so defenders with offensive automated responses have a positive duty to consider possible offensive counter-responses by the adversary, and how those can potentially interact with other automated responses of their own. There are rare situations in which defenders can reasonably conclude that a chain reaction is improbable. For example, suppose the defenders' system *S* has only one offensive automated response *R* (and the rest are defensive) and the adversary system that will be targeted by *R* is known to (almost certainly) have only counter-responses that target *S* (i.e., the source of *R*). In this case, there may be a partial escalation, but no runaway chain reaction is plausible. Such situations are likely to be the exception rather than the rule, however, as the combinatoric possibilities of offensive responses and counter-responses will typically include at least some ethically problematic chain reactions. We thus contend that there is a general presumption against offensive automated responses: they are ethically permissible only if defenders undertake the positive work of determining that the conditions for a runaway chain reaction do not exist.

A similar obligation arises from the projection bias challenge. Defenders have a positive obligation to ensure, or at least try to ensure, that their offensive automated responses will still be endorsed when they are activated in the novel cyber-warfare context. If the response is not endorsed at the later time, then it would be an act of (cyber-)warfare for which no individual or group is responsible (since responsibility inheres in intentional rather than accidental acts). Importantly, this positive obligation arises precisely because of the automated nature of the offensive response; if a human decision-maker is "in the loop," then this positive obligation largely disappears, as there will be a locus of responsibility. Even when no human intervention is possible, it may be possible to directly determine whether the decision made now (to implement the response) will likely still be endorsed when that response is triggered. First, as noted in Section 2, practices such as red team-blue team exercises might (depending on what empirical data emerge) yield decisions that accurately track the choices that decision-makers would endorse if the actual circumstances arose. Second, much of the psychological research on projection bias has aimed to find predictable patterns in people's prediction errors. As just one example, people systematically underestimate, and essentially never overestimate, the later value (to them) of objects not yet in their possession (the "endowment effect"; see, e.g., Kahneman, Knetsch, & Thaler, 1990, though List, 2003 shows that certain experiences can reduce this effect). Just as current value estimates provide a lower bound on later judgments, one might hope that current decisions about automated responses can similarly provide (partial) information about later endorsements. To our knowledge, no empirical data directly address this question, and so the viability of this response strategy is unknown.

Third, and perhaps most plausibly, one could argue that the triggering context for the offensive automated response is sufficiently similar to one's usual decision contexts, and so the decision can be trusted not to change. Projection bias arises when thinking about unfamiliar

contexts (e.g., decisions about life-sustaining medical treatments) but not for familiar contexts. For example, people are reasonably good at predicting the clothes they are likely to wear next week, precisely because it is a familiar context. Triggering cyberwarfare contexts are obviously quite unusual for most people, but might be quite common for the relevant decision-makers (e.g., individuals in the U.S. Cyber Command). Such a response to the projection bias challenge would require a detailed analysis of the workflow conditions and decision challenges for such individuals; to our knowledge, no such analysis is publicly available, and so we cannot adjudicate whether this avenue will prove viable. Given that key empirical questions are open, we contend that the positive obligation due to the projection bias challenge remains, and so there is a second presumptive barrier to adoption of offensive automated responses in cyberwarfare contexts.

4. Cyberexploitation: part of the tricky middle ground

Although the moral permissibility of the clear-cut types of automated responses is relatively straightforward (yes for defensive ones, no for offensive ones), there are other types of cyber activities whose moral permissibility is much less clear. In this section, we consider the relevance of the three challenges for the ethical permissibility of automated responses that provide, enable, or support cyberexploitation capabilities and activities. We use 'cyberexploitation' here to refer to activities in which defenders access an adversary's systems in order to gain an informational advantage, but not to (directly) disable or otherwise adversely affect those systems. Examples of actions that fall under the scope of this term include covert exfiltration of information (without damaging the system), changing entries in a database (without corrupting that database), or installing a back door to facilitate access at later times. Human operators can sometimes perform these actions deliberately, but many of them can also be implemented as part of an automated

response in certain contexts. For example, exfiltration routines can be deployed such that information, both open and hidden, is collected and returned automatically on a regular basis.

CNE occupies an unclear middle ground between the offensive and defensive extremes: it is not offensive since defenders are not seeking to damage the adversary's computer network system, infrastructure, or kinetic capabilities, but it is not defensive since defenders are directly influencing computer network systems that are not under their legitimate control. We thus see here the real value of the three challenges, as they provide a practical guide to determining the ethical permissibility of automated responses that fall into this significant middle ground. As before, we contend that the attribution challenge is not a significant one for CNE in the cyberwarfare context, as defenders frequently will have substantial, actionable information about the likely adversary. Considerable resources could of course be required to establish attribution at a high enough confidence level to justify an automated response, but governments under threat or at peril during cyberwarfare would presumably have resources to devote to establish attribution with sufficient confidence. We thus focus principally on the chain reaction and projection bias challenges.

For both of these challenges, we contend that much depends on the details of the particular CNE response, which is perhaps unsurprising given that we are dealing here with "middle ground" cases. More specifically, we distinguish between automated responses that do not directly lead to changes in the *functioning* of an adversary's systems, and those that do yield such changes (which must be non-adverse, as we are focused on cyberexploitation, not cyberattack). The former type includes actions such as exfiltration of information or installation of a back-door into the adversary's systems.[9] There is little reason to think that such actions would trigger a

[9] Installing a back-door obviously changes the adversary's systems, but the simple act of installation should not change their *functioning*, assuming that the adversary does not have

chain reaction of automated responses, precisely because they do not influence the functioning of the adversary's systems and so are much harder to detect, and respond to, automatically. Of course, such actions may well prompt a reaction from adversaries when they realize what has happened, but the chain reaction challenge focuses on sequences of actions and reactions that occur without human intervention. Having said that, adverse chain reactions of automated responses are still possible, depending on the adversary's detection capabilities. For example, an adversary may have set up an automated counter-response whenever exfiltration is detected, such as corrupting the information being sent back. Such corrupted information might then trigger an automated action by the defenders; it might even trigger an automated cyberattack. In such cases, the action-response sequences might spiral out of control, at least in the rapid timeframe of automated actions.

The projection bias challenge is not quite as easy to overcome in these cases. If no decision-maker endorses the triggering of this type of automated response *R* (at the moment of triggering), then the issue is to what extent the change (if any) to the adversary's systems constitutes an act of warfare (for which no one would be responsible). If defenders have the ability to make it "as if" *R* never occurred (e.g., destroying or never using exfiltrated information), then *R* does not constitute an act of warfare, since the adversary has suffered no harm (after the corrective action). For cases such as installing a back-door, however, defenders cannot undo the change in the adversary's systems, and so their ethical permissibility is likely to be highly context-dependent. In particular, whether this weakening of the adversary constitutes an act of warfare will depend on many particular details of the adversary's systems, the scope of the cyberwarfare activities, and more. For example, an installed back-door could afford the automatic exfiltration of information,

---

detection routines for just such back-doors. Of course, *using* such a back-door could easily change their functioning, but then it would fall into the category of either impactful CNE (if no adverse impacts) or cyberattack.

either continuously or on a schedule (regular or irregular) to avoid detection by the adversary. A human decision-maker would, however, quickly assume full moral responsibility for the CNE response, including not using the backdoor or actually disabling it. In summary, CNE automated responses that do not affect the functioning of the adversary's systems are likely to be ethically permissible, though there are important subtleties to be addressed even for these cases.

Suppose instead that the automated response *R* does induce some (non-adverse) change in the adversary's systems' functioning. For example, *R* might change information in an adversary's database (e.g., of troop strength or locations) or, more interestingly for cyber cases, change properties of the database (e.g., access privileges). Alternately, distribution lists could be modified to exclude those adversary actors who need to be "in the loop," or to include adversary actors who are not part of a particular decision making group, but who might have an operational interest in the decisions. Such changes could significantly alter the command-and-control structure of the adversary. Such an automated response would presumably be intended to (mis)lead the adversaries without them recognizing the subterfuge. In this case, it is very hard to know *a priori* whether a chain reaction of automated responses would result, as it would depend on, for example, whether the adversary's systems have programs that scan its databases for particular combinations of factors that prompt automated responses. Such programs certainly exist in other domains; for example, many stock trading systems automatically make certain trades when specific combinations occur in a central "database" (e.g., the NYSE). And it is at least possible that the adversary would have programs that could launch automated cyberattacks under some conditions (e.g., depending on information in a database about the defender's cyber-systems), though no such programs have been publicly disclosed. Unlike many of the other cases we have considered in this paper, we cannot say here whether it is likely that a chain reaction could occur or be avoided. However, the chain reaction challenge does provide a (relatively)

clear operational test for whether an automated CNE response that changes the functioning of an adversary's systems would be ethically permissible.[10]

The projection bias challenge is a bit more straightforward in this case. We assume that defenders are not willing to disclose their cyberexploitation activity to the adversary; defenders would not, for example, inform the adversary that they had changed entries in a database. The challenge thus turns on whether the change effected by *R* (e.g., the disinformation) constitutes an act of warfare for which someone should be held morally responsible. This is a particular instance of the more general question of when it is acceptable (in warfare) to deceive an adversary. With the advent of the information age, information operations, especially as they relate to denial and deception, have grown in importance in warfare. Certainly within cyberwarfare, the opportunities for, and the impact of, information operations have exploded exponentially. Full exploration of information operations within the cyber context is required to determine the boundary conditions of their ethical permissibility.

This more general issue has been the locus of substantial philosophical disagreement whose resolution depends in large part on one's account of the ethics of warfare. The most interesting space of positions holds that disinformation is sometimes, but not always, morally permissible. Accounts in this space typically focus on whether the disinformation is in "good faith" (Mattox, 1998): that is, disinformation is ethically permissible when it falls within the "rules of warfare." For example, placing inflatable tanks on Pacific islands during World War II to give the appearance of greater troop numbers is morally acceptable; having a weapons factory masquerade as a hospital is not. We suggest that similar considerations apply in the CNE domain: changing a database entry to make a training facility appear to be a battlefield hospital

[10] To be a bit more precise, the challenge only indicates whether the automated response is *im*permissible; meeting the challenge is necessary for permissibility, but not sufficient.

would be an unethical act of warfare; editing a distribution list so that information is spread (against the adversary's wishes) throughout the command-and-control structure would be acceptable. Thus, the question of whether the projection bias challenge arises for a particular automated cyberexploitation response will depend on the fine details of the response, and the broader warfare context in which it could occur.

4. Conclusion

Automated responses are an inevitable aspect of cyberwarfare, but there has not been a systematic treatment of the conditions in which they are morally permissible. We have argued that there are three substantial barriers to the moral permissibility of an automated response: the attribution, chain reaction, and projection bias problems. The first has been previously discussed, as it is a characteristic problem for *all* cyber-activities. The latter two seem not to have been discussed previously in the philosophical (or other) literature, but arguably pose a greater barrier in the cyberwarfare context. Moreover, these three challenges together provide a set of operational tests that can be used to assess the moral permissibility of a particular automated response in a specific situation. We have argued that defensive automated responses will almost always pass all three challenges, while offensive automated responses (in the cyberwarfare context) typically face a substantial positive burden in order to overcome the chain reaction and projection bias challenges. Perhaps the most interesting cases arise in the middle ground between cyberoffense and cyberdefense, such as automated cyberexploitation responses. In those situations, we see that much depends on the finer details of the response, the context, and the adversary. Importantly, however, the operationalizations of the three challenges provide a clear guide for decision-makers to assess the moral permissibility of automated responses that could potentially be implemented.

Acknowledgments

Thanks to two anonymous reviewers for their valuable comments on an earlier version of this paper. DD was partially supported by a James S. McDonnell Foundation Scholar Award. JHD's work was supported, in whole or in part, with funding from the United States Government. Any opinions, findings and conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the University of Maryland, College Park, and/or any agency or entity of the United States Government.

References

Campbell, J. P. (1968). Individual versus group problem solving in an industrial sample. *Journal of Applied Psychology, 52*, 205-210.

Conlin, M., O'Donoghue, T., & Vogelsang, T. J. (2007). Projection bias in catalog orders. *American Economic Review, 97*, 1217-1249.

Dipert, R. R. (2006). Preventive war and the epistemological dimension of the morality of war. *Journal of Military Ethics, 5*, 32-54.

Dipert, R. R. (2010). The ethics of cyberwarfare. *Journal of Military Ethics, 9*, 384-410.

Ditto, P. H., Smucker, W. D., Danks, J. H., Jacobson, J. A., Houts, R. M., Fagerlin, A., Coppola, K. M., & Gready, R. M. (2003). The stability of older adults' preferences for life-sustaining medical treatment. *Health Psychology*, **22**, 605-615.

Frankfurt, H. G. (1969). Alternate possibilities and moral responsibilities. *Journal of Philosophy, 66*, 829-839.

Guerrero, A. A. (2007). Don't know, don't kill: Moral ignorance, culpability, and caution. *Philosophical Studies, 136*, 59-97.

Hirokawa, R. Y. (1980). A comparative analysis of communication patterns within effective and ineffective decision making groups. *Communication Monographs, 47*, 312-321.

Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1990). Experimental tests of the endowment effect and the Coase theorem. *Journal of Political Economy, 98*, 1325-1348.

List, J. A. (2003). Does market experience eliminate market anomalies? *Quarterly Journal of Economics, 118*, 41-71.

Loewenstein, G., O'Donoghue, T., & Rabin, M. (2003). Projection bias in predicting future utility. *Quarterly Journal of Economics, 118*, 1209-1248.

Mattox, J. M. (1998). *The ethics of military deception*. Unpublished master's thesis. U.S. Army Command and General Staff College, Fort Leavenworth, KS.

Owens, W. A., Dam, K. W., & Lin, H. S. (Eds.). (2009). *Technology, policy, law, and ethics regarding U.S. acquisition and use of cyberattack capabilities*. Washington, D. C.: The National Academies Press.

Pereboom, D. (2000). Alternative possibilities and causal histories. *Philosophical Perspectives, 14*, 119-137.

Popper, N. (2012, 1 August). Flood of errant trades is a black eye for Wall Street. *The New York Times*.

Read, D., & Van Leeuwen, B. 1998. Predicting hunger: The effects of appetite and delay on choice. *Organizational Behavior and Human Decision Processes, 76*, 189-205.

Rosen, G. (2003). Culpability and ignorance. *Proceedings of the Aristotelian Society, 103*, 61-84.

Rowe, N. C. (2010). The ethics of cyberweapons in warfare. *International Journal of Cyberethics, 1*, 20-31.

Sanger, D. E. (2010, September 26). Iran fights malware attacking computers. *The New York Times*, p. A4.

Walzer, M. (1977). *Just and unjust wars: A moral argument with historical illustrations*. New York: Basic Books.

Wolf, S. (1987). *Responsibility, character and the emotions*. Cambridge: Cambridge University Press.

Woolley, A. W., Gerbasi, M. E., Chabris, C. F., Kosslyn, S. M., & Hackman, J. R. (2008). Bringing in the experts: How team composition and work strategy jointly shape analytic effectiveness. *Small Group Research, 39*, 352-371.

Zimmerman, M. (1997). Moral responsibility and ignorance. *Ethics, 107*, 410-426.

Author Biographies

David Danks is an Associate Professor of Philosophy & Psychology at Carnegie Mellon University (Pittsburgh, PA), as well as a Research Scientist at the Institute for Human & Machine Cognition (Pensacola, FL). His research focuses on cognitive science, principally applying techniques from machine learning and theoretical frameworks from philosophy to the study of the mind. He has published in a wide range of journals, including Journal of Philosophy, Philosophy of Science, and Psychological Review. He is currently finishing a book manuscript under contract to MIT Press exploring the use of graphical models to capture mental representations. E-mail: ddanks@cmu.edu.

Joseph H. Danks is Research Professor at the Center for Advanced Study of Language at the University of Maryland (College Park, MD). He for many years previously was Professor of Psychology and Dean of the College of Arts and Sciences at Kent State University (Kent, OH). He has published numerous journal articles, book chapters, and books on a variety of topics in psycholinguistics and cognitive psychology, including how people solve problems; how people read and comprehend written texts; how children learn to read; how skilled translators create texts in the target language; how elderly patients communicate their wishes for end-of-life medical care; how to forecast the plans and intentions of a country's leadership to develop weapons of mass destruction (WMDs); and how to develop psychological profiles of cyber adversaries. E-mail: jdanks@casl.umd.edu.