
HARVARD LAW SCHOOL
NATIONAL SECURITY JOURNAL

ONLINE ARTICLE

Perfidy in Cyberspace: The Requirement for Human Confidence

Captain Sean K. Price*

* Judge Advocate, United States Marine Corps. J.D., University of Chicago Law School, 2011; B.A., Old Dominion University, 2008. The author is presently assigned as Student, 67th Judge Advocate Officer Graduate Course, The Judge Advocate General's School, United States Army, Charlottesville, Virginia. This article was submitted in partial completion of the Master of Laws requirements of the 67th Judge Advocate Officer Graduate Course.

Table of Contents

Introduction

The United States is under attack. In the months leading up to hostilities, the enemy's intelligence agencies have identified key U.S. and allied military officials who use cloud-connected artificial pacemakers [1] or implantable cardiac defibrillators (ICD). [2] Immediately preceding offensive operations in the physical domains, the adversary's cyber force pushes malware to those officials' pacemakers, which accept it as authentic firmware updates produced by a civilian manufacturer. When the attack begins, the adversary instructs the now-infected pacemakers to malfunction.[3] The most fortunate targets require hospitalization to deactivate their pacemakers. A few senior officers[4] and civilian defense officials die.[5]

Assuming it was lawful to target these officials,[6] did the adversary violate the rule against perfidy by exploiting their pacemakers' cybersecurity vulnerabilities[7] to wound or kill them? Put briefly (and to be discussed in more detail below), the rule against perfidy prohibits inviting the adversary's confidence concerning protection under the law of armed conflict with the intention of wounding or killing him.[8] This question produced a split within the International Group of Experts who contributed to the *Tallinn Manual 2.0*[9]: the majority believed that the facts presented

---

[1] Modern pacemakers are, indeed, connected to the cloud. *See* Neta Alexander, *My Pacemaker Is Tracking Me from Inside My Body*, THE ATLANTIC (Jan. 27, 2018), https://www.theatlantic.com/technology/archive/2018/01/my-pacemaker-is-tracking-me-from-inside-my-body/551681/ [perma.cc/M4C9-4DW6] ("[E]very new pacemaker implanted in the United States is cloud-connected.").

[2] Though they are not the same, both artificial pacemakers and ICDs are implantable medical devices used to treat heart conditions. Because the difference between them is not relevant to the topic of this article, both artificial pacemakers and ICDs will henceforth be referred to simply as "pacemakers."

[3] For a discussion of how a "maliciously configured" pacemaker could be made to harm a patient, see Daniel Halperin et al., *Pacemakers and Implantable Cardiac Defibrillators: Software Radio Attacks and Zero-Power Defenses*, *in* PROCEEDINGS OF THE 2008 IEEE SYMPOSIUM ON SECURITY AND PRIVACY 129, 131 (2008).

[4] While it is theoretically possible for an individual to serve on active duty with a pacemaker, the more likely victims of such an attack would be senior civilian leaders.

[5] This scenario is based on an example of cyber perfidy given in TALLINN MANUAL 2.0 ON THE INTERNATIONAL LAW APPLICABLE TO CYBER OPERATIONS 493–94 (Michael N. Schmitt & Liis Vihul eds. 2017) [hereinafter TALLINN MANUAL 2.0].

[6] OFFICE OF GEN. COUNSEL, DEP'T OF DEF., DEPARTMENT OF DEFENSE LAW OF WAR MANUAL § 5.7.4 (2016) [hereinafter LAW OF WAR MANUAL], https://dod.defense.gov/Portals/1/Documents/pubs/DoD%20Law%20of%20War%20Manual%20-%20June%202015%20Updated%20Dec%202016.pdf [perma.cc/TBV2-HFYN] (stating that leaders, including certain civilian officials, are subject to attack).

[7] The U.S. federal government has identified pacemaker cybersecurity vulnerabilities. For example, on August 29, 2017, the Food and Drug Administration issued a recall for certain pacemakers for a firmware update to address cybersecurity vulnerabilities. *See* U.S. FOOD & DRUG ADMIN., FIRMWARE UPDATE TO ADDRESS CYBERSECURITY VULNERABILITIES IDENTIFIED IN ABBOT'S (FORMERLY ST. JUDE MEDICAL'S) IMPLANTABLE CARDIAC PACEMAKERS, https://www.fda.gov/medical-devices/safety-communications/firmware-update-address-cybersecurity-vulnerabilities-identified-abbotts-formerly-st-jude-medicals [perma.cc/E6SP-5XZW].

[8] *See* Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts (Protocol I) art. 37(1), June 8, 1977, 1125 U.N.T.S. 3 [hereinafter AP I].

[9] Though the *Manual* does not represent the official views of any state or organization, Michael N. Schmitt, *Introduction* to TALLINN MANUAL 2.0, *supra* note 5, at 1–2, it is nevertheless the most influential statement (at least in the West) of how international law currently applies to cyber warfare, *see* Michael J. Adams, *A Warning About* Tallinn 2.0 … *Whatever It Says*, LAWFARE (Jan. 4, 2017, 8:30 AM), https://www.lawfareblog.com/warning-about-tallinn-20-%E2%80%A6-whatever-it-says [perma.cc/2ZKX-XW9R] ("It is routinely referenced and relied upon by civilian and military practitioners across the globe . . . .").

would violate the rule against perfidy, while others believed that they would not, because "confidence presupposes human involvement."[10] This article argues that the latter group was correct because confidence, as that term is used in the rule against perfidy, means human trust.

Wars will have a cyber component for the foreseeable future.[11] The new, more aggressive cyber strategy of the U.S. Department of Defense (DoD) reflects this reality, calling for "defend[ing] forward to disrupt or halt malicious cyber activity at its source."[12] The DoD also expressed its intent to "employ offensive cyber capabilities and innovative concepts."[13] A clear and realistic understanding of what constitutes perfidy in cyberspace will inform the limits of that innovation.

Human confidence is—and should be—required to commit perfidy. To explain why, this article first discusses the modern definition of perfidy, the rule against it, and its function within the larger framework of lawful and unlawful deception. Second, it describes the characteristics of cyberspace that are the most challenging to the application of perfidy. Third, it explains how perfidy as currently defined and applied does not extend to the deception of cyber systems. Fourth, this article argues that the definition of perfidy should not be expanded to include the deception of cyber systems. Finally, this article anticipates some counterarguments in favor of the majority position set out in the *Tallinn Manual 2.0*.

I. Perfidy Defined

Perfidy, also sometimes called treachery,[14] "is the false claim to protections under the law of war in order to secure a military advantage."[15] The term is used both broadly to describe generally bad faith or dishonorable conduct, and more narrowly as an element of the crime of perfidy as it exists today.[16] This article is concerned with perfidy in the latter sense, which must further be distinguished from the rule against it. The crime of perfidy does not prohibit perfidy *per se*, but rather perfidy as a means of achieving a certain result: death or injury of the adversary.[17]

A. *Perfidy and the Rule Against It*

Perfidy as a concept has its origins in chivalric notions of honor and fairness between combatants.[18] Rules against perfidy were first codified during the 19th century in the Lieber Code, the Brussels Declaration, and the 1880 Oxford Manual, which prohibited "clandestine or

[10] TALLINN MANUAL 2.0, *supra* note 5, at 493–94.
[11] *See, e.g.*, P.W. SINGER & ALLAN FRIEDMAN, CYBERSECURITY AND CYBERWAR 127 (2014) (discussing Israel's hacking of Syrian air defense network in advance of a 2007 airstrike).
[12] U.S. DEP'T OF DEF., SUMMARY: DEPARTMENT OF DEFENSE CYBER STRATEGY 1 (2018) (emphasis omitted).
[13] *Id.* (emphasis omitted).
[14] *See* LAW OF WAR MANUAL, *supra* note 6, § 5.22.1.1.
[15] *Id.* § 5.22.1.
[16] *See* INT'L COMM. OF THE RED CROSS, COMMENTARY ON THE ADDITIONAL PROTOCOLS OF 8 JUNE 1977 TO THE GENEVA CONVENTIONS OF 12 AUGUST 1949 ¶ 1483, at 430 (Yves Sandoz et al. eds. 1987) [hereinafter AP I COMMENTARY] ("Literally speaking, perfidy means the breaking of faith . . . ."); LAW OF WAR MANUAL, *supra* note 6, § 5.22.1.1.
[17] MICHAEL BOTHE ET AL., NEW RULES FOR VICTIMS OF ARMED CONFLICTS: COMMENTARY ON THE TWO 1977 PROTOCOLS ADDITIONAL TO THE GENEVA CONVENTIONS OF 1949, at 234–35 (2d ed. 2013).
[18] *See id.* at 233; Sean Watts, *Law-of-War Perfidy*, 219 MIL. L. REV. 106, 171 (2014).

treacherous attempts to injure an enemy,"[19] "murder by treachery,"[20] and "perfidious, unjust, or tyrannical acts,"[21] respectively.[22] Codification continued with the Hague Regulations of 1907, which prohibit "kill[ing] or wound[ing] treacherously individuals belonging to the hostile nation or army."[23]

The modern definition of, and rule against, perfidy was codified in Article 37 of Additional Protocol I (AP I) to the Geneva Conventions of 1949:

> 1. It is prohibited to kill, injure or capture an adversary by resort to perfidy. Acts inviting the confidence of an adversary to lead him to believe that he is entitled to, or obliged to accord, protection under the rules of international law applicable in armed conflict, with intent to betray that confidence, shall constitute perfidy. . . .
>
> 2. Ruses of war are not prohibited. Such ruses are acts which are intended to mislead an adversary or to induce him to act recklessly but which infringe no rule of international law applicable in armed conflict and which are not perfidious because they do not invite the confidence of an adversary with respect to protection under the law. The following are examples of such ruses: the use of camouflage, decoys, mock operations and misinformation.[24]

AP I's formulation of the rule is generally considered to reflect customary international law,[25] though it has not entirely displaced the rule in the Hague Regulations of 1907.[26] AP I's rule against perfidy (like that in the Hague Regulations of 1907) outlaws a discrete type of deception: not perfidy itself, but rather perfidy that proximately causes the adversary's death or injury.[27] For example, approaching the enemy under a flag of truce and then attacking would violate the rule

---

[19] ADJUTANT-GEN.'S OFFICE, U.S. DEP'T OF WAR, GEN. ORDER NO. 100, INSTRUCTIONS FOR THE GOVERNMENT OF ARMIES OF THE UNITED STATES IN THE FIELD art. 101 (1863), *reprinted in* 2 THE MISCELLANEOUS WRITINGS OF FRANCIS LIEBER 245, 265 (1881).

[20] PROJECT OF AN INTERNATIONAL DECLARATION CONCERNING THE LAWS AND CUSTOMS OF WAR art. 13(b) (1874), *reprinted in* THE LAWS OF ARMED CONFLICTS 23, 24 (Dietrich Schindler & Jiri Toman eds., 4th ed. 2004).

[21] INST. OF INT'L LAW, THE LAWS OF WAR ON LAND art. 4 (1880), *reprinted in* THE LAWS OF ARMED CONFLICTS, *supra* note 20, at 29, 31.

[22] For an in-depth discussion of these provisions, see Watts, *supra* note 18, at 125–34.

[23] Convention Respecting the Laws and Customs of War on Land, Annex art. 23(b), Oct. 18, 1907, 36 Stat. 2277 [hereinafter Hague Convention IV].

[24] AP I, *supra* note 8, art. 37.

[25] *See* 1 JEAN-MARIE HENCKAERTS & LOUISE DOSWALD-BECK, INT'L COMM. OF THE RED CROSS, CUSTOMARY INTERNATIONAL HUMANITARIAN LAW 225 (2005) ("[I]t can be argued that killing, injuring or capturing by resort to perfidy is illegal under customary international law . . . ."). The United States, however, does not recognize "capture" as one of the acts prohibited by resort to perfidy. LAW OF WAR MANUAL, *supra* note 6, § 5.22.2.1. This was also the view of a majority of the International Group of Experts who contributed to the *Tallinn Manual 2.0*. *See* TALLINN MANUAL 2.0, *supra* note 5, at 492.

[26] AP I COMMENTARY, *supra* note 16, ¶ 1488, at 431; Watts, *supra* note 18, at 151. The implications of this relationship for perfidy in the cyber context are discussed *infra* Part V.B.

[27] BOTHE ET AL., *supra* note 17, at 234–35. A strict reading of the prohibition requires the attack to be successful because the text of Article 37 of AP I does not explicitly cover attempts. But commentary by the International Committee of the Red Cross (ICRC) states that "the attempted or unsuccessful act also falls under the scope of [the] prohibition." AP I COMMENTARY, *supra* note 16, ¶ 1493, at 433. This question split the International Group of Experts who contributed to the *Tallinn Manual 2.0*. *See* TALLINN MANUAL 2.0, *supra* note 5, at 493.

against perfidy.[28] Sending a small part of one's force under a flag of truce solely to delay the enemy would not, even though it satisfies AP I's definition of perfidy. Yet it would still be unlawful deception—specifically, the improper use of the flag of truce.[29] Unlike the rule against perfidy, the prohibition on improper use of a flag of truce and of other recognized emblems (e.g., the red cross) is "absolute."[30] Improper use of such an emblem is illegal, plain and simple.[31]

As this example illustrates, the rule against perfidy does not need, nor should it be expected, to do all the work of marking the boundary between lawful and unlawful deception. The definition of perfidy within the rule is narrower and more demanding than the rule set forth in the Hague Regulations of 1907, which prohibits "kill[ing] or wound[ing] treacherously."[32] Nonetheless, it is still more flexible than the rules on using recognized emblems.

Thus, perfidy's modern definition strikes a balance between flexibility and objectivity, so that it can fill the gap between recognized emblems and ruses[33] while retaining sufficient clarity to give combatants adequate notice of what deception is illegal. After all, protection under the law of war can take many forms, not all of which require a recognized emblem. Civilians, for example, do not need to wear a specific symbol to warrant protection from attack.[34] Accordingly, the rule against perfidy must address inviting the adversary's confidence with respect to legal protections that are not accompanied by a recognized emblem. Getting the balance between flexibility and objectivity right is important because a ruse's legality is defined in the negative: the deception is lawful so long as it neither is perfidious nor violates a specific rule in the law of armed conflict (e.g., improper use of the enemy's uniform).[35] Like perfidy, the concept of ruses has chivalric origins.[36] Thus, if perfidy were left in its previous form—prohibiting treacherous wounding or

[28] *See* LAW OF WAR MANUAL, *supra* note 6 § 5.22.3.
[29] Hague Convention IV, *supra* note 23, Annex art. 23(f); AP I, *supra* note 8, art. 38(1); *see also* LAW OF WAR MANUAL, *supra* note 6, § 12.4.2.1.
[30] AP I COMMENTARY, *supra* note 16, ¶ 1532, at 448.
[31] *See* BOTHE ET AL., *supra* note 17, at 235 ("[Articles 38 and 39 of AP I] prohibit the improper use of the emblems . . . concerned within their scope and thus include a prohibition of the perfidious use of these symbols even if it does not necessarily meet all of the criteria of the first sentence of Art. 37.")
[32] *See* Watts, *supra* note 18, at 108 ("[T]he twentieth century's codification of the perfidy prohibition converted a popularly and intuitively understood label for betrayal of trust or confidence into a technically bound term of art, comparatively divested of much of its customary import and broad coverage.").
[33] By "ruses," I mean "deceptions that are not prohibited by the law of war," as opposed to deceptions in warfare generally. *See* LAW OF WAR MANUAL, *supra* note 6, § 5.25.1.4.
[34] *See id.* § 4.8.1.5 (defining "civilian" as "a person who is neither part of nor associated with an armed force or group, nor otherwise engaging hostilities.").
[35] *See* AP I, *supra* note 8, art. 37(2); LAW OF WAR MANUAL, *supra* note 6, § 5.25.1. There is, however, space between the definition of ruse and the rule against perfidy for so-called "permissible perfidy": perfidious deception that neither results in the adversary's injury, death, or capture nor misuses a recognized emblem. *See* Watts, *supra* note 18, at 149–50. This gap in coverage is not relevant to this article because it argues that deception of a cyber system would not be perfidious.
[36] *See* Thomas C. Wingfield, *Chivalry in the Use of Force*, 32 U. TOL. L. REV. 111, 113 (2001) ("As strongly as the law of chivalry is woven into the fabric of the modem law of war, it remains most intact in the distinction between lawful ruses and treacherous perfidy.").

killing, without further elaboration of what "treachery" means—the distinction between perfidy and ruses would turn on subjective, and consequently unenforceable, notions of fairness.[37]

Reinforced with its present, objective standard, the rule against perfidy is better able to serve its two purposes, which are linked by a common thread of trust.[38] First, the rule seeks to preserve "the basis for restoration to peace"[39] by ensuring "reliable mediums of exchange and communication" between the parties,[40] thereby allowing resolution of the conflict "short of the complete annihilation of one belligerent by the other."[41] For example, an offer to surrender will not be trusted if the other side believes it is a trick to injure or kill its soldiers. Second, and more importantly, the rule protects persons who are not subject to attack under the law of armed conflict by assuring combatants that their respect for protected persons (e.g., civilians) may not lawfully be exploited to do them harm.[42] All this is not to say that the modern definition of perfidy is perfect, but rather that states should make sure that perfidy's role in the cyber context stays true to its role in the physical domain.

B. *The* Tallinn Manual 2.0 *and Perfidy*

The description of perfidy in the *Tallinn Manual 2.0* is, for the most part, a straightforward application of the rule set forth in AP I.[43] Two aspects of its analysis of perfidy merit mention here: its interpretation of the concept of "adversary" and its discussion of proximate cause.

The *Manual* observes that "[t]he notion of 'adversary' is sufficiently broad to encompass the situation in which the deceived person is not necessarily the person whose death or injury results from the deception."[44] It neither elaborates on this point nor cites to any authority in support of it.[45] Even so, the *Tallinn Manual 2.0*'s broad reading of "adversary" is consistent with AP I's

---

[37] *Cf.* LAW OF WAR MANUAL, *supra* note 6, § 5.21 ("The line between those deceptions that good faith permits and those that good faith prohibits may appear indistinct and has varied according to State practice.").
[38] *See* AP I COMMENTARY, *supra* note 16, ¶ 1499, at 434 ("[T]he first proposals concerned with defining the concept of perfidy were based on the concept of trust, which forms the basis of the security of international relations.").
[39] BOTHE ET AL., *supra* note 17, at 233; *see also* U.S. DEP'T OF DEF., MANUAL FOR MILITARY COMMISSIONS pt. IV, § 5(17)(c)(3), at IV-15 (2016).
[40] Watts, *supra* note 18, at 172.
[41] MANUAL FOR MILITARY COMMISSIONS, *supra* note 39, pt. IV, § 5(17)(c)(3), at IV-5.
[42] *See* BOTHE ET AL., *supra* note 17, at 233 ("Combatants, in practice, find it difficult to respect protected persons and objects if experience causes them to believe or suspect that their adversaries are abusing their claim to protection under the rules of international law applicable in armed conflict in order to achieve a military advantage."); Watts, *supra* note 18, at 171 ("Civilians, the wounded, and those offering surrender or truce enjoy more reliable protection when soldiers are confident that their forbearance in attacking these persons will not be betrayed or used against them.").
[43] *See* TALLINN MANUAL 2.0, *supra* note 5, at 491 ("In the conduct of hostilities involving cyber operations, it is prohibited to kill or injure an adversary by resort to perfidy.").
[44] *Id.* at 492.
[45] *See id.* Indeed, the implementation of the rule against perfidy in both the U.S. Military Commissions Act of 2009 and the Elements of Crimes under the Rome Statute of the International Criminal Court apparently require the person or persons deceived and the person or persons killed, wounded, or captured to be the same. *See* 10 U.S.C. § 950t(17) (2018) (specifying that victims of perfidy are "such person or persons" whose confidence or belief had been invited); INTERNATIONAL CRIMINAL COURT, ELEMENTS OF CRIMES art. 8(2)(b)(xi), at 24 (2011), https://www.icc-cpi.int/NR/rdonlyres/336923D8-A6AD-40EC-AD7B-45BF9DE73D56/0/ElementsOfCrimesEng.pdf [perma.cc/QA94-E2PU] (same).

definition and is essential to the application of perfidy in cyberspace, as shown by its discussion of proximate cause.

To violate the rule against perfidy, "the perfidious act must be the proximate cause of the death or injury."[46] To illustrate, the *Manual* gave this example: a military unit sends an email to the adversary "indicating an intention to surrender some days later at a specific location," but then ambushes the adversary unit sent to accept the surrender.[47] Those responsible for sending the email have violated the rule against perfidy, regardless of whether any member of the ambushed unit knew of the email's existence.[48] The email proximately caused the ambush by deceiving a person with the authority to dispatch the unit; it need not have deceived any of the victims.

So far, so good. However, the International Group of Experts split on the scope of an adversary's confidence, specifically whether it includes the confidence of a cyber system.[49] Put another way, can the deception of a computer system be imputed to the adversary? To explain why it would be odd to think so, a discussion of some of the peculiar aspects of cyberspace is warranted.

II. Aspects of Cyberspace

It is by now axiomatic that international law, including the law of armed conflict, applies to cyberspace.[50] The DoD recognizes cyberspace as its own domain.[51] And cyberspace is firmly anchored in the real world,[52] the ambitions of information freedom enthusiasts[53] notwithstanding. Every piece of information in cyberspace can be tied to a physical place—that is, the server or computer on which it resides—and is thus subject to the jurisdiction of one or more states.[54] Nevertheless, cyberspace has two properties that frustrate the application of law-of-armed-conflict concepts in general, and perfidy in particular: cyberspace is (1) inherently artificial and (2)

---

[46] TALLINN MANUAL 2.0, *supra* note 5, at 492; *accord* BOTHE ET AL., *supra* note 17, at 235; Watts, *supra* note 18, at 154.

[47] TALLINN MANUAL 2.0, *supra* note 5, at 493. This example is also intended to show that temporal proximity is not a requirement for proximate causation. *See id.*

[48] *See id.* at 492–93.

[49] *See id.* at 493–94. The *Manual* defines a cyber or computer system as "[o]ne or more interconnected computers with associated software and peripheral devices." *Id.* at 564.

[50] *See* Rep. of the Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security, ¶ 28(b), U.N. Doc. A/70/174 (July 22, 2015) ("Existing obligations under international law are applicable to State use of [information and communications technologies]."); Gary D. Solis, *Cyber Warfare*, MIL. L. REV., Spring 2014, at 1, 1–2 (affirming that "existing [laws of armed conflict] apply to cyber issues," *id.* at 1).

[51] *See* JOINT CHIEFS OF STAFF, JOINT PUB. 3-12, CYBERSPACE OPERATIONS I-1 (8 June 2018) [hereinafter JOINT PUB. 3-12] ("[T]he [DOD] is responsible for defending the US homeland and US interests from attack, including attacks that may occur in cyberspace." (citation omitted)).

[52] *See* SINGER & FRIEDMAN, *supra* note 11, at 182 ("[E]very node of the network, every router, every switch is within the sovereign borders of a nation-state . . . or travels on a submarine cable or satellite connection owned by a company that is incorporated in a sovereign nation-state . . . ." (citation omitted)).

[53] *See, e.g.*, John Perry Barlow, *A Declaration of the Independence of Cyberspace*, ELECTRONIC FRONTIER FOUND. (Feb. 8, 1996), https://www.eff.org/cyberspace-independence [perma.cc/TJ39-J4PC] ("Governments of the Industrial World . . . [y]ou have no sovereignty where we gather [in cyberspace, that is].").

[54] *See* TALLINN MANUAL 2.0, *supra* note 5, at 13 ("A State enjoys sovereign authority with regard to the cyber infrastructure, persons, and cyber activities located within its territory, subject to its international legal obligations.").

thoroughly civilian in terms of its architecture, means, and users. These two properties significantly complicate the analysis required to distinguish between ruses and perfidy.

A. *Cyberspace Is Inherently Artificial*

There is no such thing as a *natural* environment or background in cyberspace. It is "the realm of computer networks,"[55] which are by definition *artificial*. To be sure, it has a human element because people use computer networks. This is reflected in DoD's cyberspace layer model, which describes cyberspace "in terms of three interrelated layers: physical network, logical network, and cyber-persona."[56] Put simply, the physical and logical network layers are the hardware (e.g., a desktop computer) and the code transcending it (e.g., a website that "exists on multiple servers in multiple locations").[57] Cyber-personas "consist[] of network or [information technology] user accounts, whether human or automated, and their relationships to one another."[58] For instance, a person's social media account is part of the cyber-persona layer; it thereby serves as a direct link between the person and the logical network layer.[59] But, of course, cyber-personas are themselves artificial. For instance, "[o]ne individual may create and maintain multiple cyber-personas . . . which may vary in the degree to which they are factually accurate."[60]

B. *Cyberspace Is Generally Civilian*

Cyberspace is the culmination of a general trend in technology and warfare that has been progressively complicating the application of the principle of distinction: the merging of civilian and military technology and infrastructure. As Professor Michael Schmitt has observed: "[I]t is becoming ever more difficult to determine when an object, and the facility that makes it, is military."[61] Cyberspace epitomizes this problem because "[t]he private sector owns and operates over ninety percent of all of the networks and infrastructure of cyberspace."[62] Consequently, something like "98 percent of U.S. government communications, including classified communications, travel over civilian-owned-and-operated networks."[63] It is no surprise, then, that "nearly all cyber operations occur on, in, or through civilian cyberspace infrastructure."[64]

C. *Camouflage in Cyberspace*

---

[55] SINGER & FRIEDMAN, *supra* note 11, at 13.
[56] JOINT PUB. 3-12, *supra* note 51, at I-2.
[57] *Id.* at I-3 to I-4.
[58] *Id.* at I-4.
[59] *See id.*
[60] *Id.* For an extreme example of the maintenance of multiple cyberpersons, see, for example, P.W. SINGER & EMERSON T. BROOKING, LIKEWAR: THE WEAPONIZATION OF SOCIAL MEDIA 138 (2018) (discussing Russian "botnet" of at least 60,000 Twitter accounts).
[61] Michael N. Schmitt, *The Principle of Discrimination in 21st Century Warfare*, 2 YALE HUM. RTS. & DEV. L.J. 143, 159 (1999).
[62] U.S. DEP'T OF DEF., THE DEPARTMENT OF DEFENSE CYBER STRATEGY 5 (2015).
[63] SINGER & FRIEDMAN, *supra* note 11, at 196 (citation omitted).
[64] Gary P. Corn & Peter P. Pascucci, *The Law of Armed Conflict Implications of Covered or Concealed Operations: Perfidy, Ruses, and the Principle of Passive Distinction*, *in* THE IMPACT OF EMERGING TECHNOLOGIES ON THE LAW OF ARMED CONFLICT 273, 277 (Michael N. Schmitt et al. eds. 2019).

These two aspects of cyberspace complicate the perfidy analysis by blurring the line between perfidy and camouflage, which Article 37 of AP I lists as an example of a ruse.[65] The purpose of camouflage is to blend into the background to avoid detection even under direct observation (as opposed to taking cover behind an object that obscures observation).[66] In cyberspace, the only background to blend into is man-made and mostly civilian. This, by itself, is not necessarily a problem. The law of armed conflict does not prohibit blending into a man-made, civilian background, such as in a city.[67] But "[t]he feigning of civilian, non-combatant status" is perfidious.[68] Thus, "[a] combatant . . . can use camouflage and make himself virtually invisible against a natural or man-made background, but he may not feign a civilian status and hide amongst a crowd."[69]

Cyberspace seldom allows for this distinction between blending into the background and hiding in a crowd, for in cyberspace's logical network layer, they are one and the same. Thus, "[f]requently, concealment in cyberspace requires, in effect, hiding in plain 'technical' sight to evade identification and attribution."[70] Hiding this way is essential for cyberspace operations to avoid detection by adversary personnel or programs.[71] The stakes are especially high in the cyber context, because the adversary's detection of a particular cyberspace capability will not only render that capability useless against that adversary, but also allow the adversary to replicate the capability for its own uses.[72] For example, Stuxnet—a cyberweapon jointly developed by the United States and Israel to attack Iranian nuclear centrifuges[73]—"may have taken the combined efforts of a team of experts almost a year to build," but it only took a few weeks after its discovery for a blogger to post an online how-to guide to building it.[74] Shortly thereafter, variations of Stuxnet (such as the "son of Stuxnet") began to appear "in the wild."[75]

These operational realities virtually eliminate any obligation of passive distinction with respect to cyberspace operations. Passive distinction requires parties to "distinguish or separate [their] military forces and war-making activities from members of the civilian population to the maximum extent feasible."[76] There is no extent to which a party can feasibly distinguish a cyberspace capability from surrounding, generally civilian, code.[77] For example, owing to the U.S. government's near-total reliance on civilian infrastructure for its communications, that same

[65] AP I, *supra* note 8, art. 37(2).
[66] *See* AP I COMMENTARY, *supra* note 16, ¶ 1507, at 438.
[67] *See id.*; Corn & Pascucci, *supra* note 64, at 292.
[68] AP I, *supra* note 8, art. 37(1)(c).
[69] AP I COMMENTARY, *supra* note 16, ¶ 1507, at 438.
[70] Corn & Pascucci, *supra* note 64, at 292.
[71] *See generally* SINGER & FRIEDMAN, *supra* note 11, at 61–62 (discussing antivirus programs and firewalls).
[72] *See* JOINT PUB. 3-12, *supra* note 51, I-12.
[73] SINGER & FRIEDMAN, *supra* note 11, at 114–18.
[74] *Id.* at 158.
[75] *Id.* at 159.
[76] LAW OF WAR MANUAL, *supra* note 6, § 2.5.3.
[77] *See* Corn & Pascucci, *supra* note 64, at 298.

infrastructure is the principal conduit through which the United States' adversaries may conduct cyber operations against it.[78]

III. Back to Perfidy

What, then, does this mean for the application of perfidy in cyberspace? In short, it means that, just like in the physical domains, the deception of a human being is a necessary—and desirable—component of perfidy. Expanding the concept of "adversary" to include the adversary's cyber systems and that of "confidence" to include the perception of those systems would not only go beyond what is required by customary international law, but would also unduly restrict cyberspace capabilities. This Part first explains why the current definition of perfidy requires the deception of a human being. It then argues that perfidy should not be expanded to include inviting the confidence of a cyber system.

A. *Human Confidence Is Required to Commit Perfidy*

Perfidy is premised on the deception of a human for two reasons. First, the notion of an adversary implies humanity. Second, confidence—that is, trust—is uniquely human. The reasoning here is essentially textual, and the relevant text is that of Article 37 of AP I.

As discussed in the *Tallinn Manual 2.0*, the word "adversary" in Article 37 is general enough to include multiple people.[79] Depending on the context in which it is used, it can refer to a single soldier, a unit, an entire fighting force, its commander-in-chief, or anything in between.[80] But it does not refer to things, such as tanks, planes, rifles, computers, and, yes, cyber systems. An adversary is a party to a conflict, or, typically, a member or members of a party's armed forces.[81]

More fundamentally, wars are fought by people, not things. The law of armed conflict reflects this principle by distinguishing between persons and objects.[82] Admittedly, the term "object" does not apply neatly to cyber systems because, on the majority view, objects must be "visible and tangible."[83] Thus, data, including software, are not objects under the law of armed conflict.[84] A minority of the International Group of Experts for the *Tallinn Manual 2.0* believed

---

[78] For its part, DoD recognizes this problem: "Many of DOD's critical functions and operations rely on contracted commercial assets, including Internet service providers (ISPs) and global supply chains, over which DOD and its forces have no direct authority." JOINT PUB. 3-12, *supra* note 51, I-12 to I-13.

[79] *See* TALLINN MANUAL 2.0, *supra* note 5, at 492.

[80] *Cf.* JEAN DE PREUX, INT'L COMM. OF THE RED CROSS, 3 THE GENEVA CONVENTIONS OF 12 AUGUST 1949: COMMENTARY 50 (Jean S. Pictet ed., A.P. de Heney trans., 1960) ("[T]he term 'enemy' covers *any adversary* during an 'armed conflict which may arise between two or more of the High Contracting Parties' . . . ." (emphasis added) (first quoting Geneva Convention Relative to the Treatment of Prisoners of War art. 4(A), Aug. 12, 1949, 6 U.S.T. 3316; and then quoting *id.* art. 2)).

[81] *See* LAW OF WAR MANUAL, *supra* note 6, § 4.2 ("The law of war has recognized that the population of an enemy State is generally divided into two classes: the armed forces and the civilian population.").

[82] *Compare, e.g.*, AP I, *supra* note 8, art. 51 (protection of civilian population), *with, e.g.*, *id.* art. 52 (protection of civilian objects).

[83] AP I COMMENTARY, *supra* note 16, ¶¶ 2007–08, at 633–34.

[84] *See* TALLINN MANUAL 2.0, *supra* note 5, at 437.

that sufficiently important data should be considered objects because it would be absurd to allow the deletion of data with impunity, no matter the severity of the consequences.[85]

Regardless, under either view, a cyber system can only be an object. The only question is how much of the system is considered an object. While the rule against perfidy itself does not use the word "person," it refers to the adversary in personal terms by prohibiting the adversary's killing, injury, or capture by resort to perfidy.[86] Objects cannot be killed or injured, and when taken, they are "seized," not captured.[87]

Where enforced, the rule against perfidy reflects this understanding. The United States has made perfidy triable by military commission.[88] The rule prohibits "inviting the confidence or belief of one or more *persons*" concerning protection under the law of armed conflict and using "that confidence or belief in killing, injuring, or capturing such *person or persons*."[89] Similarly, the Elements of Crimes under the Rome Statute of the International Criminal Court require "[t]he perpetrator [to have] invited the confidence or belief of one or more *persons*."[90]

The current definition of perfidy also requires the deception of a human by specifying that the deception must "invit[e] the confidence of an adversary to lead him to believe" that he or someone else is entitled to legal protection.[91] Computers do not have beliefs. They store and process information. They do not make judgments concerning legal protection. Even if they did, current law would not recognize their decisions on protection under the law of armed conflict. That responsibility would remain with the cognizant commander.[92] Thus, the minority within the International Group of Experts for the *Tallinn Manual 2.0* correctly took "the position that the notion of confidence presupposes human involvement."[93]

B. *Why Human Confidence Should Be Required*

If the current definition of perfidy does not include inviting the confidence of a cyber system, the natural follow-up question is whether it should. There have been, after all, calls for new, overarching international agreements on cyberspace (a "Digital Geneva Convention," for

[85] *See id.*
[86] AP I, *supra* note 8, art. 37(1).
[87] *See, e.g.*, LAW OF WAR MANUAL, *supra* note 6, § 5.17.2 ("Enemy property may not be *seized* or destroyed unless imperatively demanded by the necessities of war." (emphasis added)).
[88] *See* 10 U.S.C. § 950t(17).
[89] *Id.* (emphasis added); *accord* MANUAL FOR MILITARY COMMISSIONS, *supra* note 39, pt. IV, § 5(17)(b), at IV-14 (elements of the offense).
[90] ELEMENTS OF CRIMES, *supra* note 45, art. 8(2)(b)(xi), at 24 (emphasis added).
[91] AP I, *supra* note 8, art. 37(1).
[92] *See* LAW OF WAR MANUAL, *supra* note 6, § 6.5.9.3 ("[I]t is persons who must comply with the law of war.").
[93] TALLINN MANUAL 2.0, *supra* note 5, at 494.

instance).[94] Nothing has come of these calls thus far,[95] but customary international law can change even without formal agreements.[96] So, if a critical mass of states followed the majority view of perfidy expressed in the *Tallinn Manual 2.0* out of a sense of legal obligation, it would become the law.[97] Should they? No, for four reasons. First, the cyber systems themselves do not need the protection of the rule against perfidy. Second, existing prohibitions against unlawful deception adequately regulate the use of cyberspace capabilities. Third, such a rule would be both theoretically unsound and impractical to apply. Finally, deceiving a cyber system does not implicate the same interests that the rule against perfidy is meant to protect.

First, as discussed above, since a cyber system is, at most, an object, the law of armed conflict protects the system based on its "nature, location, purpose or use."[98] Cyber systems thus do not require the protection of the rule against perfidy as they are adequately protected by the rules concerning objects. Of course, one might argue that perfidy and the rule against it are more concerned with protecting people than things.

The point, then, of expanding the definition of perfidy to include inviting the confidence of a cyber system would be to protect certain persons, rather than the cyber systems themselves. This argument fails because the rules concerning objects are already structured to incidentally protect people who must not be made the object of attack. For example, combatants are generally required to respect and protect the cyber systems of medical units, thereby incidentally protecting medical personnel and their patients (who are presumably *hors de combat*).[99] Any cyber system that is not a military objective would similarly (though not as strictly) be protected from attack.[100] Behind these rules, the proportionality principle protects civilians and other protected persons by prohibiting attacks expected to produce excessive collateral damage.[101]

---

[94] *See, e.g.*, Brad Smith, President, Microsoft Corp., Keynote Address at the RSA Conference 2017: The Need for a Digital Geneva Convention (Feb. 14, 2017), https://blogs.microsoft.com/wp-content/uploads/2017/03/Transcript-of-Brad-Smiths-Keynote-Address-at-the-RSA-Conference-2017.pdf [perma.cc/D9FP-8YNN] (proposing a Digital Geneva Convention "that will call on the world's governments to pledge that they will not engage in cyberattacks on the private sector, that they will not target civilian infrastructure").

[95] *See, e.g.*, Arun M. Sukumar, *The UN GGE Failed. Is International Law in Cyberspace Doomed as Well?*, LAWFARE (July 4, 2017, 1:51 PM), https://www.lawfareblog.com/un-gge-failed-international-law-cyberspace-doomed-well [perma.cc/C22C-N5B4] (noting the U.N. Group of Governmental Experts failed to reach consensus on "how international law applies to the use of Information and Communication Technologies (ICTs) by states").

[96] *See* LAW OF WAR MANUAL, *supra* note 6, § 1.8 ("Customary international law is an unwritten form of law in the sense that it is not created through a written agreement by States.").

[97] *See id.* ("*Customary international law* results from a general and consistent practice of States that is followed by them from a sense of legal obligation (*opinio juris*).").

[98] AP I, *supra* note 8, art. 52(2); *accord* LAW OF WAR MANUAL, *supra* note 6, § 5.6.3 (quoting *id.*).

[99] *See* TALLINN MANUAL 2.0, *supra* note 5, at 515 ("Computers, computer networks, and data that form an integral part of the operations or administration of medical units and transports must be respected and protected . . . .").

[100] *See id.* at 435–45 ("Cyber infrastructure may only be made the object of attack if it qualifies as a military objective." *Id.* at 434); *see also* AP I, *supra* note 8, art. 52(1) ("Civilian objects shall not be the object of attack or of reprisals. Civilian objects are all objects which are not military objectives . . . .").

[101] *See* LAW OF WAR MANUAL, *supra* note 6, § 5.10; *see also* TALLINN MANUAL 2.0, *supra* note 5, at 470 ("A cyber attack that may be expected to cause incidental loss of civilian life, injury to civilians, damage to civilian objects, or a combination thereof, which would be excessive in relation to the concrete and direct military advantage anticipated is prohibited.").

Second, existing rules that distinguish between lawful and unlawful deception, including the rule against perfidy in its current form, already regulate the use of cyberspace capabilities to deceive humans. The *Tallinn Manual 2.0*'s example of emailing the adversary falsely claiming that a military unit will surrender is illustrative.[102] Betraying the adversary's confidence that the unit will surrender by attacking violates the rule against perfidy. Had no person been deceived by the email, the adversary would not have sent the unit that got ambushed. It does not matter whether the cyber system itself was deceived.

Other rules of unlawful deception also apply, such as the improper use of recognized emblems.[103] To be sure, they do not translate neatly to the cyber context. Consider, for example, the question whether the protected indicator itself must be used (e.g., the red cross) or if an email purporting to originate from someone with an "@icrc.org"[104] email address is sufficient to violate the law of armed conflict.[105] This question split the *Tallinn Manual 2.0*'s International Group of Experts.[106] However, expanding the definition of perfidy to include inviting the confidence of a cyber system would not solve this problem because, as discussed below, it is not clear what it means to invite the confidence of a cyber system.[107] Instead, the more sensible reform would be to expand the definition of recognized emblems to include "apparently authoritative indication[s]" of the protected status each emblem is intended to represent, such as the @icrc.org domain.[108]

Perfidy and the rule against it are an essential part of the legal framework regulating deception, and means and methods generally, in armed conflict. That said, perfidy should not be expected to play a larger role in cyberspace than it does in the physical domains. Indeed, because of cyberspace's entirely artificial and principally civilian character, perfidy will probably play a much smaller role within it vis-à-vis the principles of distinction and proportionality. A target's legal susceptibility to attack (e.g., by virtue of membership in the armed forces) and proportionality considerations are more important in regulating cyber means and methods than the rule against perfidy is.

Third, expanding the definition of perfidy to include inviting the confidence of cyber systems would be unwise both theoretically and practically. Theoretically, what exactly would it mean for a cyber system to extend its confidence based on some piece of code (e.g., malware) purporting to have protected status? Take the pacemaker scenario from the beginning of this article: is it accurate to say that a pacemaker's software accepted malware because it believed the malware had civilian, non-combatant status or was otherwise protected under the law of armed conflict? No, because the software simply did what it was programmed to do—it did not actually have a choice. That is, its response to the malware was determined by its programming. The malware

---

[102] *See* TALLINN MANUAL 2.0, *supra* note 5, at 493.
[103] *See id.* at 496–504 (prohibiting improper use of protective indicators, United Nations emblem, enemy indicators, or neutral indicators).
[104] Official ICRC email addresses end with "@icrc.org." *Fraudulent E-mails and Websites*, ICRC, https://www.icrc.org/en/faq/fraudulent-emails-and-websites [perma.cc/ZK6M-EUZK] (last visited Jan. 21, 2020).
[105] *See* TALLINN MANUAL 2.0, *supra* note 5, at 498.
[106] *See id.*
[107] At least with respect to the possibility that the email address will deceive a person. It would bear on the scenario used in the *Tallinn Manual 2.0* where "an email message spoofed to originate from the 'icrc.org' domain . . . bypass[es] the enemy's data filters and deliver[s] a piece of malware to the military network." *Id.*
[108] *Id.* at 499.

fulfilled a certain set of conditions the pacemaker's software was programmed to look for, and, on seeing those conditions fulfilled, the pacemaker's software mechanistically accepted the malware.

This theoretical problem—that there is no such thing as a cyber system's confidence—would give rise to practical difficulties. For example, a change in the definition of perfidy might incentivize designers to program their software to describe its actions in legal terms, such as by having log files say things like "Update accepted due to civilian, non-combatant status." But that would do nothing to change the nature of the processes governing the cyber system's behavior. By its nature, a cyber system does not have confidence, so the question whether a cyber system's confidence has been invited cannot, in fact, be answered.

In addition to the fundamental problem that a cyber system does not act or refrain from acting based on a piece of code's protected status, there is the question of what exactly civilian, non-combatant status looks like in cyberspace. The drafting of Article 37 of AP I foreshadowed this problem. Originally, "[t]he [International Committee of the Red Cross's] draft listed 'the disguising of combatants in civilian clothing'" as an example of perfidy.[109] It was changed to "feigning of civilian, non-combatant status"[110] because specifying disguise in civilian clothing as perfidy "might be misused to punish some combatants who would be entitled to prisoner of war status."[111] The point was to prohibit combatants from fooling their opponents into believing they are civilians, not to prohibit the use of civilian clothing as such. This distinction does not translate well to cyberspace. When all of cyberspace is manmade, and virtually all of it is civilian or civilian-made, what is the difference between malware merely wearing civilian clothing and malware feigning civilian status?

There is no meaningful way to make such a distinction from the perspective of a cyber system without taking most cyberspace capabilities off the table. What, after all, does it mean for a cyber system to perceive a piece of code as having civilian status rather than fail to identify its true nature because it is camouflaged? To be effective, "hidden malware . . . mimics the innocuous, usually civilian, objects or lines of code that surround it."[112] The importance of cyberspace capabilities to the future of warfare is widely acknowledged. The world's foremost military powers have incorporated cyberspace into their plans, and invested in cyberspace capabilities accordingly. At the same time, the U.S. military is dependent on civilian cyber infrastructure.[113] Together, this means that a rule prohibiting or substantially limiting the camouflage of cyberspace capabilities against a civilian background would simply not be followed.

Finally, it must be asked what interests an expanded definition of perfidy would vindicate. It would serve neither of the rule against perfidy's purposes: maintaining "the basis for restoration to peace"[114] and protecting the classes of persons who are not subject to attack under the law of

---

[109] BOTHE ET AL., *supra* note 17, at 236 (citation omitted).
[110] AP I, *supra* note 8, art. 37(1)(c).
[111] BOTHE ET AL., *supra* note 17, at 236; *see also* AP I, *supra* note 8, art. 44(3) (granting prisoner-of-war status to combatants who "owing to the nature of the hostilities . . . cannot" distinguish themselves from the civilian population if they satisfy certain requirements).
[112] Watts, *supra* note 18, at 167.
[113] *See* JOINT PUB. 3-12, *supra* note 51, at I-13 (acknowledging DoD's "[d]ependency on commercial Internet providers").
[114] BOTHE ET AL., *supra* note 17, at 233.

armed conflict.[115] An expanded definition of perfidy would not preserve the potential for trust *between adversaries*. It would instead enhance the degree of trust adversaries can have (or, probably more accurately, believe they can have) *in their cyber systems*. This is not a worthy goal for the law of armed conflict.

In an analogous context, the law offers no protection. The use of "enemy codes, passwords, and countersigns" is permissible,[116] although it is generally unlawful to use "enemy flags, insignia, and military uniforms" in combat.[117] The two forms of deception deserve different treatment because "military forces are expected to take measures to guard against the use of their codes, passwords, and countersigns by the enemy."[118] The same cannot be said of flags, insignia, and uniforms because they, like recognized emblems in general, cannot perform their function when concealed or constantly changed.

The law of armed conflict is not intended to protect the adversary or the means by which it wages war.[119] It is not, in other words, a shield behind which military forces are supposedly able to rest assured that their systems are uncompromised. So, just as the law of armed conflict generally offers no protection for codes, passwords, or countersigns, it should not generally protect military cyber systems. This is what the concept of perfidy would tend to do if expanded to include the confidence of cyber systems. Military forces should be expected to take measures to guard the integrity of their cyber systems. By retaining perfidy's requirement that a person be deceived, adversaries could trust the reliability of communications that invite their confidence about legal protection while still having to exercise diligence regarding what their cyber systems are doing behind the scenes.

IV. Some Counterarguments

There are three important counterarguments to the position outlined in this article. The first is that a strict reading of the rule against perfidy draws an arbitrary distinction between functionally equivalent results. The second is that an expanded definition of perfidy can be read into Article 23 of the Hague Regulations of 1907. Finally, there is the argument that not expanding the definition of perfidy will discourage the outsourcing of distinction decisions to systems that may, in the future, be better than people at making those decisions.

A. *Functional Equivalence*

The strongest counterargument is the functional equivalence between deceiving people to harm them and tricking a cyber system to harm them in the same way. Take the scenario from the beginning of this article—where the officials' pacemakers automatically accept the malware disguised as updates (call this scenario *A*)—and suppose instead that the model of pacemakers in

---

[115] *See supra* note 42.
[116] LAW OF WAR MANUAL, *supra* note 6, § 5.23.1.5; *see also* BOTHE ET AL., *supra* note 17, at 246.
[117] LAW OF WAR MANUAL, *supra* note 6, § 5.23.1.5; *see also* AP I, *supra* note 8, art. 39(2) ("It is prohibited to make use of the flags or military emblems, insignia or uniforms of adverse Parties while engaging in attacks or in order to shield, favour, protect or impede military operations.").
[118] LAW OF WAR MANUAL, *supra* note 6, § 5.23.1.5.
[119] *See id.* § 1.3.4 (listing purposes of the law of war).

question requires owners to affirmatively accept updates through a pop-up on their smartphones (call this scenario *B*). Why does it make sense for the malware to be considered perfidy in scenario *B* but not in scenario *A*? Suppose further that owners of the pacemaker can determine whether they must affirmatively accept updates, but the malware can change the settings to accept updates automatically (call this scenario *C*). In this case, a human has not been deceived because the otherwise perfidious means of cyber-attack has deprived the person of the opportunity to be—or avoid being—deceived.

Under all three sets of facts, the malware is functionally equivalent: it is masquerading as an authentic firmware update from a civilian manufacturer in order to harm or kill the pacemaker's host. The consequences of the malware's success are the same. Yet, on the view of perfidy advanced by this article, the malware's sender has not violated the rule against perfidy in scenario *A* or *C*. Only in scenario *B*, when an affirmative human action was required to accept the malware, has the element of inviting the adversary's confidence been fulfilled. It seems arbitrary to deem legal an action that harms or kills people without their ever being aware of the means of their injury or demise, but to call the exact same action illegal if it happens to require the victims to be fooled about the source of the attack.

The simple, but unsatisfying, response to that objection is that the law often draws arbitrary distinctions. Sometimes this is because of conflicting values or principles in the drafting process. Other times, legal distinctions do not seem arbitrary until new circumstances arise. Or, as is the case with the rule against perfidy, both of these things are true. The rule does not prohibit perfidy per se, thereby leaving open the possibility of permissible perfidy.[120] And the internet did not exist when the modern rule was codified in AP I. Consequently, it would be surprising if Article 37 translated perfectly to cyberspace.

This article goes one step further than simply pointing out that perfidy in its current form does not apply to scenario *A* or *C*: it argues that it should not apply. This is because the gravamen of the offense is unlawful deception, not injury or death.[121] It is the targets' combatant status that renders them subject to injury or death. Absent deception, the death or injury of a combatant does not reduce the credibility of inter-personal communication. Thus, prohibiting the conduct described in scenarios *A* and *C* does not advance the purposes of the rule against perfidy.

B. *What About the Hague Regulations of 1907?*

Recall that Article 37 of AP I did not displace the already-existing prohibition codified in Article 23 of the Hague Regulations of 1907,[122] which prohibits "kill[ing] or wound[ing] treacherously individuals belonging to the hostile nation or army."[123] The word "treacherously" is sufficiently broad to encompass the deception of cyber systems. The majority view of the *Tallinn*

[120] *See* Watts, *supra* note 18, at 149–50.
[121] *See* John C. Dehn, *Permissible Perfidy?: Analysing the Colombian Hostage Rescue, the Capture of Rebel Leaders and the World's Reaction*, 6 J. INT'L CRIM. JUST. 627, 644 (2008) ("The possible results of prohibited perfidy—meaning death, injury or capture—are not forbidden between combatants in armed conflict. It is the use of bad faith to obtain those results that is wrongful." (footnote omitted)).
[122] AP I COMMENTARY, *supra* note 16, ¶ 1488, at 431.
[123] Hague Convention IV, *supra* note 23, Annex art. 23(b).

Manual 2.0's International Group of Experts could, therefore, be tied to an already-existing legal provision.

The problem with this argument is that while the word "treacherously" is indeed broad, it is not specific enough to be legally determinate. Legal reasoning alone cannot answer the question whether deceiving a cyber system can be treacherous. Reference to norms and customs is necessary to determine the content of the word. This lack of specificity is the problem that led to the creation of the more precise formula set out in Article 37 of AP I.[124] Moreover, states have not employed cyberspace capabilities in armed conflict enough to have formed norms and customs[125] that supply widely accepted meaning to the word "treachery" in the cyber context. Consequently, a debate about the legal definition of "treachery" would collapse into a conversation about whether it *should* include the deception of cyber systems.

C. *Outsourcing Distinction Decisions*

The counterargument that is the most speculative, but perhaps most consequential in the long run, is that a human-centric definition of perfidy will impede the development of technologies capable of making distinction decisions better than people can. It should be expected that artificial intelligence (AI) will be developed to the point of being able to distinguish between lawful and unlawful targets (however that is defined in the AI's code).[126] Professor Schmitt gives the example of "an autonomous anti-personnel weapon system designed for employment in urban areas" with "sufficient sensor and artificial intelligence capability to distinguish [between civilians and combatants]."[127]

There are a number of reasons why such a system might be better at making distinction decisions. For instance, it would presumably not be subject to factors (like fatigue and fear) that can adversely affect human judgment. The employment of such a system would also reduce casualties for the side employing it.[128] But if combatants are permitted to exploit the programming designed to make a system comply with the principle of distinction—for example, by affixing to their clothing "adversarial stickers" designed to be interpreted by the autonomous weapon as a red cross but imperceptible to humans[129]—parties may refrain from using such systems, despite their superiority in making distinction decisions. Or worse, parties may relax the standards for distinction in the systems' programming.

A full discussion of the principle of distinction as it relates to autonomous weapon systems is outside the scope of this article. Suffice it to say that modifying the definition of perfidy is not

---

[124] *See* AP I COMMENTARY, *supra* note 16, ¶ 1489, at 431–32 (explaining "the inadequate wording of the Hague Regulations" forms part of the background "for providing a general definition of perfidy for the first time").
[125] *See* Solis, *supra* note 50, at 2.
[126] *See* Christopher M. Ford, *Autonomous Weapons and International Law*, 69 S.C. L. REV. 413, 434–439 (2017) (discussing how an autonomous weapon might distinguish between persons).
[127] Michael N. Schmitt, *Autonomous Weapon Systems and International Humanitarian Law: A Reply to the Critics*, HARV. L. SCH. NAT'L SECURITY J. (Feb. 5, 2013), https://harvardnsj.org/2013/02/autonomous-weapon-systems-and-international-humanitarian-law-a-reply-to-the-critics/ [perma.cc/9VH7-NWZQ].
[128] *See* Ford, *supra* note 126, at 432 ("A single combatant could control dozens of autonomous weapons systems, which could replace hundreds or thousands of combatants.").
[129] *Cf.* Ryan Calo et al., *Is Tricking a Robot Hacking?* 7 (Univ. of Wash. Sch. of Law, Legal Studies Research Paper No. 2018-05, 2018), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3150530 [perma.cc/Q8A4-W66G] (giving the example of an "adversarial sticker" that caused a self-driving car to misidentify a traffic sign).

the most effective means of accommodating the development of such systems within the law of armed conflict. The rule against perfidy is a relatively flexible catch-all, but human confidence is essential to its role in filling the gap between ruses and other rules of unlawful deception. It is those other rules that should be modified, or supplemented, in order to accommodate the development of distinction-capable technologies. To use the "adversarial sticker" example, the rule against improper use of recognized emblems could be expanded to include images intended to appear as such an emblem to a machine.

Conclusion

Perfidy requires human confidence, and should continue to do so. The broader view taken by a majority of the *Tallinn Manual 2.0*'s International Group of Experts neither is supported by the text of Article 37 of AP I, nor would work to preserve the possibility of trust between adversaries. Human confidence is central to that purpose.

Cyber systems are not, at least for the time being, capable of confidence in the sense that term is used in the definition of perfidy. Accordingly, expanding perfidy to include the confidence of a cyber system would entail legal analysis about something that does not exist. In addition, the cyber domain's inherently artificial and principally civilian character would tend to make compliance with such a rule impossible. Only by retaining the requirement for human confidence may perfidy continue to perform its proper function in cyberspace.