Humans, Agents and International Humanitarian Law: Dilemmas in Target Discrimination

Marten Zwanenburg, Hans Boddens Hosang, Niek Wijngaards

[1,2] Netherlands Ministry of Defense, Directorate of Legal Affairs, P.O. Box 20701, 2500 ES, The Hague, The Netherlands
{M.Zwanenburg, JFR.Boddens.Hosang}@mindef.nl
[3] Thales Research & Technology Netherlands / DECIS Lab, P.O. Box 90, 2600 AB, Delft, The Netherlands
Niek.Wijngaards@icis.decis.nl

Abstract. The severity of the (potential) consequences of the use of agents in the military domain imperatively necessitates research on the application of the law of armed conflict, which provides 'constraints' on 'human' warfare. In this paper, one of the main principles of that law, *target discrimination*, is used to highlight legal and ethical dilemmas involving the use of autonomous agents. The incident involving the USS Vincennes warship and the destruction of a commercial airliner is used to illustrate these dilemmas.

1 Introduction

Research in the area of agent technology, and in the related area of autonomous systems, progresses at a steady pace. Current developments [5] include Internet agents capable of providing information, negotiating and closing contracts (e.g. to buy a book), and robots e.g. capable of autonomously playing soccer (http://www.robocup.org/). Legal implications of the use of agent technology are being researched in the civilian domain, with emphasis on, for example, whether agents are legal personae [12] or can close contracts [14]. Consequences of the actions of agents in the civil domain are usually to a large extent reversible, e.g. by declaring a contract void, returning goods, etc. However, when agents (and/or autonomous systems) are employed in the military domain, consequences of actions may not be reversible at all: consider human suffering and the loss of human life as well as destruction of important natural resources and objects belonging to the cultural heritage of mankind.

In armed conflict situations, the law of armed conflict or International Humanitarian Law (IHL) applies. This law governs the protection of non-combatants in warfare situations, and aims to limit the effects of war on those not (directly) involved therein. IHL sets out the legal framework for the actions of commanders in the field and defines the methods and means of warfare which can legitimately be employed. An analysis of agent technology from the particular perspective of IHL is

important for a number of reasons. These include the fact that if the use of agent technology leads to violations of IHL the users may be held criminally responsible for war crimes. It may be noted that the Statute of the International Criminal Court provides that someone who facilitates the commission of a war crime by providing the means for its commission, may also be criminally responsible. Another reason is that the military is regarded as an important potential market for agent technology. IHL however requires the military to determine whether the use a weapon or means of warfare it intends to acquire would be prohibited by international law. Obviously, if the outcome is that the use would lead to violations of international law, there will be a strong disincentive to acquire the weapon or means of warfare.

In this paper, one of the main principles of IHL, *target discrimination*, is used to highlight legal and ethical dilemmas involving the use of autonomous agents. First, the involvement of agent technology in the military domain is briefly described in Section 2. A case study in which a US warship accidentally destroyed a commercial airliner is introduced in Section 3. In section 4, IHL is introduced, after which problems with the application of the IHL principle of discrimination arising from the use of agent technology are identified in Section 5. Section 6 concludes with a brief discussion.

2 Agent Technology and the Military Domain

Agent technology appears to have the active interest of armed forces in a number of states[1],[2] [13]. Agent technology is one of many terms used to describe new developments in warfare, also referred to as 'Revolution in Military Affairs' (RMA). Two similar approaches are used to 'flatten' the rigid military command and control hierarchy: "Network Centric Warfare" (NCW) by the USA, and "Network Enabled Capability" (NEC) by NATO and the United Kingdom. The UK states "NEC is about the coherent integration of sensors, decision-makers and weapon systems along with support capabilities." NCW is characterized as "the ability of geographically dispersed forces (consisting of entities) to create a high level of shared battle space awareness that can be exploited via self-synchronization and other network-centric operations to achieve commanders" intent [1]. Agent technology is an enabling technology for NEC/NCW, notably the planning and execution of military operations. For example, a key concept in NCW is the effective linking among entities in the battle space, which requires a robust, high-performance information infrastructure, or 'infostructure', that provides all elements of the warfighting enterprise with access to high-quality information services.

In our view, the concepts of agents and autonomous systems overlap: An agent is considered to be an autonomous entity, which interacts with other agents and objects, services, etc. in its environment. Whether an agent is cooperative, able to migrate, solely resides on the Internet or has a hardware body (robot), etc., is irrelevant to our discussion. Our perspective does not conflict with definitions of agents and agency such as found in common literature (e.g., see [5]). To facilitate the discussion of

[1] See http://www.darpa.mil.
[2] See http://www.jfcom.mil/about/fact_alpha.htm

agent-related dilemmas, three roles of agents are distinguished. Although agents in general may combine multiple roles, this distinction facilitates determining possible consequences of an agent's involvement:

- *Information provision*: agents that filter and aggregate information (e.g. involving 'data fusion') can be used by combat commanders to address information overload and time-criticality and relevance of information. These agents are similar to information providing agents on the Internet for civilian purposes, such as finding the best price for a consumer good.
- *Decision making*: agents that generate goals, plans and schedules can be used by a commander as more or less autonomous tactical decision support tools.
- *Action execution*: agents that are able to autonomously execute actions, including control over weapons and other equipment can be used by commanders to perform missions. Although current Unmanned Combat Aerial Vehicles (UCAVs) are remotely controlled by human operators[3], next generation UCAVs are expected to be fully autonomous [2].

3 The USS Vincennes incident

On July 3, 1988, the USS Vincennes, a Ticonderoga Class cruiser equipped with the highly sophisticated Aegis system, was engaged in a skirmish with small Iranian speedboats in the Persian Gulf. The vessel was in the area to protect American interests and shipping from the effects of the protracted war between Iraq and Iran (the "first" Gulf War). While fighting the speedboats and in the midst of the confusion and chaos resulting from maneuvering the ship and engaging multiple targets at once, a contact thought to be an Iranian F-14 aircraft was seen to be inbound to the Vincennes. Assumed to be on a descending course straight for the ship, the air contact was tagged hostile and ultimately engaged and destroyed. The contact later turned out to be a civilian Airbus type aircraft, Iran Air flight 655 from Bandar Abbas to Dubai, carrying roughly 290 passengers. There were no survivors.

There is considerable controversy surrounding various aspects of the incident, including why the Vincennes was in the location she was in and the necessity to be engaged in the surface battle with the speedboats. It is not the intention of this paper, however, to repeat the investigations or to pass judgment. Instead, the incident is used here as a platform for conjecture. If the three types of agent technology discussed above were introduced (more extensively than they were) in the Aegis system, what would the outcome have been and would the result be acceptable in the context of IHL?

The system was designed to provide extensive battle space management and enable area defense against multiple air, surface and sub-surface targets. The level of autonomy with which the system could engage targets could be increased or decreased according to the threat level and circumstances. As a blue water warfare

[3] For example, in November 2002 a Predator unmanned aerial vehicle fired a missile killing six suspected terrorists in Yemen; W. Pincus, U.S. Strike Kills Six in Al Qaeda, *The Washington Post*, 5 November 2002, at A01.

system, it was an ideal concept. In the crowded context of the Persian Gulf, however, with a confusing mix and number of military and civilian vessels and aircraft of various nationalities, it was less than ideal. Even less ideal was that the main displays showed only certain types of information (contacts and their tracks), while information about specific contacts (altitude, speed, etc.) had to be called up on separate screens. These screens did not show rate of change, however, which had to be calculated by the operators themselves. Finally, the Identification Friend or Foe (IFF) system and sensors relayed the information about contacts in the sensor gate set by the operators. The sensor gate had to be reset manually or the system would continue to relay data from the last position at which it was set, instead of the data for the actual contact at its new location. This was one of the main causes of misidentifying flight 655 (on a civilian IFF Mode) in mid-air with an actual Iranian F-14 (on a military IFF Mode) on the runway at Bandar Abbas airport.

4 International Humanitarian Law

International Humanitarian Law is also known as the Law of War or the Law of Armed Conflict (LOAC). IHL seeks to limit the effects of armed conflict. It protects persons who are not or are no longer participating in the hostilities, and restricts the means and methods of warfare which the participants in hostilities may employ. This branch of law must be distinguished from the branch of law which regulates the legality of resorting to the use of armed force. IHL is not concerned with whether a particular conflict is legal or not, but with limiting the effects once a conflict exists. It is consequently only applicable during *armed conflict*. This is the principal difference with human rights law, which is applicable also in peace-time.

IHL is strongly related to, but different from, other fields of law. One difference is that many IHL obligations are universal, i.e. they apply to (nearly) all states. In contrast, the civil law of contracts for example varies from state to state. A unique characteristic of IHL is that it must be applied in the fog of war and that it regularly involves life-or-death decisions. This means that the factors which contribute to respect or violation of the law are different from those which obtain in other fields of law. Finally, during an armed conflict IHL to a large extent sets aside other law.

The main treaties in the field of IHL are the four 1949 Geneva Conventions and the two 1977 Additional Protocols to the 1949 Conventions. Unfortunately, there is no single authoritative definition of ‘armed conflict’. Although the drafters of the Geneva Conventions appear to have had an interpretation of ‘armed conflict’ as ‘physical confrontation’ in mind, the underlying purposes of IHL must be taken into account. As Schmitt states, given advances in methods and means of warfare, it is not sufficient to apply an actor-based threshold for application of humanitarian law; instead, a consequence-based one is more appropriate ([10], p. 374). In any event, it appears that agents are usually employed in the context of a physical confrontation, as the incident involving the USS Vincennes illustrates.

A number of principles are generally recognized as fundamental principles of IHL. One of these is the principle of proportionality. This principle recognizes the inevitability of incidental damage in the attack of legitimate targets (i.e. according to

the principle of discrimination discussed below). Such incidental damage, however, may not be excessive or disproportionate in relation to the concrete and direct military advantage expected to be gained. Another is the principle of discrimination, which requires that a distinction be made between combatants and military objectives on the one hand, and civilians and civilian objects on the other hand. The latter principle will be used in this article to illustrate some important consequences which the use of agents may have for the application of IHL.

5 The Principle of Discrimination

The principle of discrimination basically states that combatants and military objectives may legitimately be attacked, and that it is prohibited to attack non-combatants and civilian objects. Its codification in Article 48 of Additional Protocol I[4] is complemented by a number of other provisions in the Protocol which relate to a number of consequences of the distinction, including Article 51(2) which prohibits attack of civilians and Article 52 which defines military objectives. Of central importance to the application of the principle of discrimination to agent technology is the interpretation of the words "effective contribution to military action" and "definite military advantage" in Article 52.[5] The ICRC Commentary states that a "definite military advantage" must be distinguished from potential or indeterminate advantages [8]. This interpretation, as well as the interpretation used by many states, rules out attacks against targets which are relatively far removed from actual military operations [6]. Other commentators and states, including the USA, adopt a broader interpretation including "war-sustaining" objects within the category of military objectives.

Commanders are required to apply IHL during armed conflict. When a commander must decide on whether to launch an attack, it is expected that the commander makes decisions on the basis of all of the information available to him at the time, notwithstanding that a tribunal that must judge his decision after the fact may have more or better information at hand. In particular, the principle of target discrimination must take into account the necessity for decision-making in the fog of war [7]. Agent technology in the role of information providing agents can significantly affect the quantity and quality of the information available to the military commander, or at any rate present the available information better. A commander can then more precisely determine whether non-combatants and civilian objects will be at risk from a planned attack. However, the application of agent technology has two important consequences. First, the latitude extended to the commander to make mistakes

[4] "In order to ensure respect for and protection of the civilian population and civilian objects, the Parties to the conflict shall at all times distinguish between the civilian population and combatants and between civilian objects and military objectives and accordingly shall direct their operations only against military objectives."

[5] "Those objects which by their nature, location, purpose or use make an effective contribution to military action and whose total or partial destruction, capture or neutralization, in the circumstances ruling at the time, offers a definite military advantage."

decreases, as the assumed limitations in the information at his disposal decrease as well. Second, a commander may rely increasingly on indirect, filtered, information, which may not contain information relevant to IHL issues.

An important finding in the investigations of the USS Vincennes incident was the concept of scenario fulfillment, combined with a (false) trust in the technology and the information it was providing. Put simply, the operators expected the situation to be one of attack and consequently interpreted the information to fit their expectations. The Aegis system filtered information in the sense that it displayed only certain information on the main displays. The system was designed for extensive blue water ('on the ocean') engagements in which civilian interlopers were not considered a realistic scenario element. Furthermore, the system was capable of tracking hundreds of contacts at once and providing extensive details on each would have overloaded the operators. Nowadays, since the attacks on the USA of September 11, 2001, the difference between an innocent civilian airliner and a guided weapon has faded significantly. Had the Vincennes been fitted with an information filtering agent, it would have been an interesting challenge to choose what information needed to be filtered out and what information would need to be displayed to the operators. Would, given these factors, the IFF mode be relevant or would instead track behavior, airspeed and angle of attack (ascent or descent) be more relevant to display? The information required to distinguish between a truly civilian object and a valid military target varies according to the circumstances and cannot be quantified or qualified uniformly to fit every possible scenario.

In short, can a commander trust that the information has been filtered in such a way that he has all the information he needs to make decisions compatible with IHL? This issue becomes even more important in multinational operations, when the same agents may filter information for commanders of different nationalities. The sending states of these commanders may not be bound by the same obligations of IHL. In this regard it is important to note that some of the principles of IHL are set forth only in Additional Protocol I, which has not been ratified by a number of important military powers, notably the United States. This fact is of importance, as present-day agent technology introduced into multinational operations is most likely to be American.

Whether agents will become capable of correctly distinguishing civilian and military targets is an open question, as the issue is also challenging to humans (e.g., consider terrorism). Although, in general, a civilian loses the protection of IHL if and for such time as he takes a direct part in hostilities[6], it is controversial how extensively "taking a direct part in hostilities" should be interpreted. A number of interpretations are advanced [4], which employ a relatively high threshold which must be crossed before active participation is accepted, in contrast to other interpretations [3]. An in-depth study of direct participation is beyond the scope of this article, but is of importance to agent technology: can agents or their human masters become military targets as well? Often civilian contractors or non-military government agencies are employed, for example by providing support for agent technology operations of armed forces, or operation of agent technology by civilians. Depending on the extent of the definition of direct participation, e.g. a civilian agent-technology expert loses IHL protection, simply by maintaining the platform of a communications

[6] Article 51 (3) Additional Protocol I.

agent without which the combat commander cannot function. In addition, the distinction between military and civilian objects becomes more blurred, e.g. as telecommunications networks and technological equipment are of a dual-use nature and may host both civilian and military (mobile) agents.

In order to make it possible to apply the principle of discrimination, combatants are obliged to distinguish themselves from non-combatants. IHL does not require participants to wear a uniform, it is sufficient that at least they carry their weapons openly during each military engagement, and during such time as they are visible to the adversary while they are engaged in a military deployment preceding the launching of an attack in which they are to participate.[7] However, it is infeasible to apply these requirements to computers and agents. Other distinguishing characteristics need to be defined, such as hosting military agents only on agent platforms (i.e., computers) with designated military IP addresses [10]. Mobile military agents, however, need other solutions; measures and policies need to be defined and implemented to constrain military agents to military agent platforms. Even then it may be unavoidable that civilian agents providing information to military agents become military targets.

Another aspect of the application of the principle of discrimination is agent autonomy: Agents operate with a certain independence from human operators, such as in unmanned combat aerial vehicles. In the Vincennes incident, the Aegis system in autonomous mode might not have made the decision to engage the civilian aircraft, given the information which was available to the system but not picked up by the human operators. But would the system have engaged the aircrafts hijacked on September 11, 2001? How would such a system make a decision capable of withstanding legal review and scrutiny which balances civilian lives on the one hand and civilian lives combined with a possible military objective on the other hand? Clearly such a decision cannot be based solely on numerical comparisons (killing a few civilians is better than killing a lot of civilians). The law dictates that the evaluation be based on a comparison between the anticipated loss of civilian lives and the military advantage anticipated from the attack. Furthermore, such an evaluation only becomes relevant after the initial determination that the actual target is a valid military objective. That in turn requires an evaluation of the information available about the target, including complex estimations of the effective military contribution of dual-use objects or goods under the specific circumstances prevailing at the time.

As Schmitt states, despite human frailties, most would agree that there are advantages to having real "eyes on target." Thus, in much the same way that technology counteracts human error, human observation mitigates technological shortcomings [11].

[7] Article 44 (3) Additional Protocol I.

6 Discussion

Russel [9] states that "people cannot be removed from interacting with computing systems for anything but the simplest, most predictable of tasks" and that "tasks that […] critically depend on external world state are often not well suited for autonomic computing." While these statements are based on an analysis of effective user interaction with autonomic computing platforms, they are certainly no less valid in the context of an analysis of autonomic agent technology in relation to the complexities of command decisions within the framework of IHL. While the Vincennes was equipped with (rudimentary) versions of the agent technology under discussion in this paper, it is clear that more extensive implementation of autonomous processes would have merely resulted in shifting the dilemmas sketched above from the shoulders of the commanding officer and crew of the USS Vincennes to the shoulders of the system designers and engineers. It follows, therefore, that although operating a complex weapons platform in context of the modern battlefield is impossible without the aid of extensive automation and, ideally, autonomous technology, targeting and engagement decisions provide case and environment-specific legal dilemmas that may *defy* resolution by autonomous agents on the basis of pre-determined "blanket" criteria.

Acknowledgements

Wijngaards' contribution to the research reported here is part of the Interactive Collaborative Information Systems (ICIS) project, supported by the Dutch Ministry of Economic Affairs, grant nr: BSIK03024, and hosted by the DECIS Lab (http://www.decis.nl).

References

1. Alberts, D.S., Garstka, J.J., Stein, F.P.: *Network Centric Warfare, Developing and Leveraging Information Superiority* (1999), 2nd rev. ed.
2. Brezinski, M.: The Unmanned Army, *The New York Times*, (2003), April 20th.
3. Guillory, M.E.: Civilianizing the Force: Is the United States Crossing the Rubicon? *Air Force Law Review*, **51** (2001), pp. 111
4. Semanza, v. Laurent (Prosecutor): International Criminal Tribunal for Rwanda (ICTR), *Judgment* (2003), Case No. ICTR-97-20-T, Tr. Ch. III, 15 May 2003, para. 364.
5. Luck, M., McBurney, P. & Preist, C.: *Agent Technology: Enabling Next Generation Computing: A Roadmap for Agent-Based Computing*. AgentLink, ISBN 0854 327886 (2003), at http://www.agentlink.org/roadmap/
6. Meyrowitz, H.: Le Bombardement Strategique d'apres le Protocole I aux Conventions de Geneve, *Zeitschrift für ausländisches öffentliches Recht und Volkerrecht*, **41** (1981) 1-68.
7. Parks, W. Hays: Linebacker and the Law of War, *Air University Review*, **34**:2 (1983) 2-30.
8. Pilloud, C., Sandoz, Y., Swinarski, C., Zimmermann, B.: *Commentary on the Additional Protocols of 8 June 1977 to the Geneva Conventions of 12 August 1949*, (1987).

9. Russel, D.M. et al.: Dealing with Ghosts: Managing the User Experience of Autonomic Computing, *IBM Systems Journal*, **42**:1 (2003) 177-188.
10. Schmitt, M.N.: Wired Warfare: Computer Network Attack and Jus in Bello, *International Review of the Red Cross*, **84** (2002) 365-399.
11. Schmitt, M.N.: The Impact of High- and Low-Tech Warfare on the Principle of Distinction, at http://www.ihlresearch.org/ihl/feature.php?a=45 (last visited 2 January 2005).
12. Sorge, C., Bergfelder, M.: Signatures by Electronic Agents: A Legal Perspective, In: Cevenini, C. (editor), *The Law and Electronic Agents: Proceedings of the LEA 04 workshop* (2004), pp. 141-153.
13. van Zijderveld, E.J.A, Maris, M.G., Brongers, D.M.: Robots in het Veld, *Militaire Spectator*, **174** (2005) 14-22, (in Dutch).
14. Weitzenböck, E.M.: Electronic Agents and the Formation of Contracts, *International Journal of Law and Information Technology*, **9**:3 (2001) 204-234.