Defensive Autonomous Weapon Systems, Humanitarian Intervention, and the Golem of Prague

A Modern Golem

Much of the existing literature seems to treat the ethical problems posed by autonomous weapon systems (AWS) as novel issues contemporaneous to technological advancements in artificial intelligence and robotics.[1] However, Jewish folklore provides a historical account of autonomous systems with combat applications, offering strong stances on the ethics of both their development and deployment. The Golem of Prague, a mythical artificial man, acts autonomously to protect Prague's Jews from outside aggression.

To our knowledge, purely defensive autonomous weapon systems (DAWS) have not been seriously considered in the current philosophical literature. Much of the literature focuses on the dangers AWS poses to human rights but fails to consider how they could instead protect them. While the Golem narratives primarily serve as inspiration for DAWS, they also provide valuable insights into the challenges associated with these systems. Just as science fiction can shape our perspective of current technologies, a critical analysis of this folklore can shed light on issues surrounding AWS.

This paper examines how the Golem narratives can contribute to the philosophical literature on the ethics of AWS and help resolve practical issues related to humanitarian intervention. After introducing the historical stories of golems, we define DAWS through a list of axioms explicitly inspired by the Golem of Prague that

constrain such a system's behavior. Next, we use a series of hypothetical cases to examine how DAWS's use of force is grounded in other-defense and proportional to the threat posed to a vulnerable community. After that, we examine the relative merits of DAWS versus traditional humanitarian intervention or arming and training vulnerable communities under the Responsibility to Protect (R2P) doctrine. Ultimately, we conclude that a Golem-inspired DAWS may offer a more effective approach to protecting these communities from atrocities.

The Golem Narratives

Jewish folklore is rich with stories of Golems, humanoid constructions made of mud or clay and endowed with autonomy by a rabbi. The exact method of animation varies between narratives but usually involves the rabbi writing words on the Golem's flesh or placing a tablet bearing the name of God in the Golem's head or mouth.[2] Although Golem stories differ widely in detail, they commonly feature artificial beings lacking human souls and possessing various abilities created by an especially learned rabbi. The history of the Golem folklore is as rich as the myths themselves, with the exact history,[3] authenticity,[4] and meaning[5] of the narratives remaining the subject of debate. While we do not weigh in on these questions, we aim to provide a concise history of the Golem folklore. We focus on how these narratives have evolved and what sets Rosenberg's Golem apart from earlier versions.

An early report of a Golem appears in the Babylonian Talmud around 1,500 years ago.[6] In this account, the sage Rava creates a speechless (artificial) man who is promptly destroyed. There are similar short stories from the Middle Ages of scholars

creating Golems that lack basic capabilities. These tales generally center around the early Kabbalistic text *Sefer Yetzirah* and contain instructions for creating a Golem, along with recommendations against doing so on the grounds of idolatry.[7]

In the 1500s-1800s, legends of Golems take root among German and Polish Jews both as an oral tradition[8] and in post-1600 manuscripts.[9] These stories are more fleshed out than the barebones accounts of the Middle Ages. Many of these early narratives give a thematically consistent account of rabbi Elijah Ba'al Shem creating a Golem in Chelm to serve as a helper around the temple. Likely referencing the legend of R. Elijah, Jakob Grimm (of the Brothers Grimm) wrote a similar short piece in 1808 about an anonymous Polish rabbi.[10] In these stories, the Golem grows larger and stronger each day. In most stories, however, it grows too strong and must be destroyed. In the process, the rabbi is usually injured or killed.[11]

Following Grimm, the Golem narratives rapidly develop into more elaborate folktales. The typical setting of the legend shifts from Rabbi Elijah Ba'al Shem of Chelm to Rabbi Judah Lowe ben Bezalel, referred to as the Maharal of Prague. In these narratives, the Golem of Prague wreaks havoc after growing too strong, sometimes destroying the temple and even attacking Jews. During this period, the Golem also appears in non-Jewish romantic novels by non-Jews like Arnim and Auerbach.[12]

In 1909, Rabbi Yehudah Yudl Rosenberg published *The Golem and the Wondrous Deeds of the Maharal of Prague, further altering the content of the Golem narratives.*[13] Written in Warsaw during a period of increasing antisemitic violence in nearby Russia, this collection of stories reimagines the reasons for creating Golems,

transforming the Golem of Prague from a mere domestic helper into a heroic protector of the Jewish people.[14] This extremely influential substrate of the legend gives the Golem a more positive image in the subsequent literature.[15] To this day, Rosenberg's Golem remains a powerful influence on popular culture. Rosenberg's book sparked renewed literary interest in stories of Golems, with several notable films, novels, and theatrical productions about Golems throughout the early 1900s.[16] Fittingly for this paper, Golems influenced Capek's play *RUR*, the etymological source of the word "robot" in English.[17] Such works and their influence cement the Golem's enduring legacy in contemporary science fiction.

Rosenberg's Golem

While the entire corpus of Golems could be philosophically relevant to the AWS literature, it is Rosenberg's Golem that inspires our vision of autonomous weapon systems as defenders of vulnerable communities. The reader would be wrong to think of our lengthy discussion of these narratives and their rich history as an unnecessary digression from the point of philosophical interest. Stories like Rosenberg's give thinkers a touchpoint for evaluating intuitions and forming beliefs. Humanitarian organizations, anti-AWS NGOs, and even some prominent AWS scholars evoke and sometimes directly reference James Cameron's *Terminator* in their names, statements, and publications.[18] It is rhetorically powerful, especially in public discourse, to provide an alternative fictional reference for a benevolent AWS when making a positive case for its use as a defender of vulnerable communities. Rosenberg provides precisely the example we need: Yossele the Golem.

Rosenberg's novel *The Golem and the Wondrous Deeds of the Maharal of Prague* is set in a time of turmoil. Prague's Jews are facing persecution from the Christian community, and the state is either unwilling or unable to help. In Rosenberg's story, many Christians commit "blood libel," falsely accusing Jews of killing Christians, usually children, to use their blood in rituals.[19] Some even try to frame Jews for murders, resulting in antisemitic violence. While the state does not levy these accusations, it does not step in to prevent the framings.[20]

In response to this situation, the Maharal of Prague creates a Golem out of loam and clay. By performing incantations while running "circuits" around the figure, the Maharal brings the Golem to life.[21] Although the Golem cannot speak, he immediately understands and obeys all of the Maharal's orders. The Maharal, wanting to keep the community unaware of the Golem's existence, gives him the name "Yossele" and disguises him as a new temple assistant. The Maharal gives Yossele a primary objective:

> "Know that we created you out of the dust of the earth to guard the Jews from all harm and from all the ills and troubles they suffer at the hands of their enemies and oppressors … No matter where I send you, you will obey each one of my commands, even enter into a blazing fire, immerse yourself in deep water, or leap from a tower until you complete the task I have given you."[22]

Obediently protecting Jews is effectively Yossele's "default state." Envisioning Rosenberg's Golem as an AWS, this primary objective underscores two of Yossele's essential features. First, Yossele is created for a fundamentally defensive purpose. His

mission is to protect the Jews of Prague, not to attack the Christians or exact revenge. Second, he must follow the Maharal's orders and defer to him.

To fulfill his mission, Yossele wanders Prague's streets disguised as a gentile, looking for anyone who might attempt to frame Jews for ritual murder. The force Yossele uses is always proportional to the force used against him. He never kills anyone and only trades blows when attacked first. When Yossele catches a perpetrator, he forcibly takes them to the authorities for arrest.

By requiring Yossele to take criminals to the state instead of exacting revenge, Rosenberg implies that the appropriate use of a Golem is as a defensive instrument. Yossele uses minimal force and avoids violence whenever possible. The Maharal could have instructed Yossele to use his immense strength for offensive purposes such as revenge killings of Christians. However, he chose to use the Golem solely to protect Prague's Jews and explicitly forbade revenge attacks.

Outside of the standing order to protect the Jews, Yossele is used for other tasks. Occasionally, he is ordered to assist with domestic tasks such as catching fish or fetching water. He is also commanded toward actions that would be difficult or dangerous for humans. For example, Yossele is used to investigate a cellar that will imminently collapse and search for a missing person for days without rest.

Yossele has several capabilities that make him suitable for his missions. First, he has immense strength and can easily overpower human adversaries. Second, he possesses a magic amulet that can turn him invisible to evade detection. Third, he has exceptionally high visual acuity, allowing him to distinguish between people more effectively than humans. Fourth, Yossele can perform missions requiring extreme

endurance and focus. Finally, he can quickly perform complex linguistic tasks at superhuman levels, like immediately solving a complicated anagram that the Maharal cannot.

Yoselle's abilities demonstrate his value as a defensive system. He can distinguish friend from foe, outfight human aggressors, carry out long, grueling tasks, camouflage himself, operate continuously, and perform complicated computations. These abilities are essential for carrying out his role as a defender of the Jewish people.

Despite not being human, Yossele can acquire and apply new information in real time. In one story, a young Jew is held captive in a secret chamber of a church cloister. Yossele learns how to work the locking mechanisms, breaks in, and successfully plans and executes an escape. This feat is notable because Yossele orchestrates and carries out the operational details without human assistance.

Eventually, Yossele and the Maharal achieve peace for Prague's Jews. The king issues an edict banning all trials against Jews for ritual murder. Rumors of Yossele even cause a decrease in blood libel in neighboring countries. Without a present need for a defender, the Maharal decommissions Yossele, reverting him back to clay and loam. His body is kept in the synagogue in case Prague's Jews need protection again.

At the end of the book, Rosenberg delineates the moral differences between a Golem and a human. He first notes that Golems do not have good or evil impulses but rather follow orders and act toward self-preservation. Further distinguishing the Golem from humans, Yossele lacks any sexual desires. Additionally, Golems are not bound by religious law, cannot be counted in a minyan, do not need to perform mitzvahs, and do

not possess divine souls. Since humans do have divine souls, it follows that the life of a human is intrinsically worth more than that of a Golem. This reading is supported by the Maharal putting the Golem into dangerous situations instead of humans.

Rosenberg's narrative provides a model for understanding the complex interplay between a persecuted minority, an oppressive majority, and a fair but indifferent state. Notably, the state itself was not an active persecutor of the Jews but instead allowed their persecution to take place. By exonerating Jews of high-profile ritual killings, Yossele's actions brought about a change in the law that ended much of the Jewish persecution. The mere threat of the Golem served as a deterrent to antisemites inside Prague and around Europe. Once the blood libel accusation stopped, Yossele was decommissioned, again demonstrating the Maharal's commitment to using the Golem only in a defensive manner.

Golems and Autonomous Weapon Systems

We now summarize some of the salient features of Golems:

- **Commanded:** Golems are given tasks to complete by the rabbis who create them. These tasks range from simple chores and errands to complex, abstract directives requiring the identification and execution of instrumental subgoals. Yossele, for instance, was always directed by the Maharal or his wife. Otherwise, he would follow his "base directive" to protect Prague's Jews.
- **Autonomous:** Golems act independently to complete their pre-specified goals. The rabbi does not actively control golems; instead, they choose their own methods and actions for their work. Notably, there were long stretches of

minimal contact between Yossele and the Maharal, leaving Yossele to act entirely independently.

- **Independent Learning and Growth:** Golems can learn, reassess situations, and change their behavior to complete their tasks better. They are not locked into a pre-specified set of potential behaviors. As described earlier, while rescuing a hostage, Yossele learned to operate a complex locking mechanism from watching the kidnappers operate it.
- **Lacking moral worth:** Despite looking and acting like humans, Golems canonically do not have human souls. Although autonomous, they do not require moral consideration. This does not mean that people do not care about the well-being of Golems, but merely that it is pro tanto morally permissible to destroy one. In the narratives, rabbis decommission Golems without remorse.

These features illustrate homology between the Golems of Jewish folklore and contemporary visions of AWS. Both Golems and AWS are given goals to carry out and act toward their completion. Executing these tasks may involve making decisions along the way and learning in the field. Neither are given moral consideration. All else being equal, losing a Golem or AWS is better than losing a human.

We caution the reader against taking the Golem as merely a framing device. The Golem narratives are legitimate philosophical sources that ground our arguments in literature written by scholarly rabbis. Ethicists regularly use fictional works like *The Adventures of Huckleberry Finn* and *Anna Karenina* to illuminate contemporary normative issues.[23] Influential Jewish scholars wrote Golem folklore about the powers

and actions of great rabbis. These folklore stories are more akin to proverbs intended to impart cultural wisdom (e.g., Jesus 'parables and Aesop's fables) than they are to urban legends (e.g., Bigfoot and the Loveland Frog). Nevertheless, we believe our forthcoming arguments in this paper stand on their own.

Another potential methodological critique is that focusing exclusively on Rosenberg's Golem overlooks the broader narrative tradition where Golems, as autonomous creations, sometimes go awry. Although Yossele never injured any innocent people, skeptics may question whether the earlier tales of Golems serve as a warning against the development and use of autonomous systems. Our reading of these stories is that they demonstrate the need for *meaningful human control* over AWS. Further, our paper's place in the AWS literature can be seen as parallel to Rosenberg's place in Golem folklore. Just as Rosenberg took the Golem, expanded its ability to conform to human values, and reimagined it as a defender of the Jewish people, we take traditionally conceived AWS, assume it conforms to certain axioms, and present it as a protector of vulnerable communities.

The Golem Axioms

To apply Rosenberg's vision of the Golem to AWS, we need to detail the type of system we are considering. Like Yossele, our conception of an autonomous weapon system is developed and deployed to protect vulnerable communities. Vulnerable communities lack internal self-defense mechanisms, encompassing states or minority populations within states that are unable or unwilling to protect them. To that end, we

present a collection of what we call the Golem Axioms, inspired by Rosenberg's narrative.

A few brief caveats are in order. First, the Golem Axioms are not to be confused with hard-coded rules like Asimov's *Three Laws of Robotics* or Arkin's *Ethical Governor*.[24] That is, we are not *delineating internal rules* but instead *constraining the system's external behavior*. This distinction merits emphasis because scholars have highlighted how unpredictable behavior poses a significant obstacle to the ethical deployment of AWS.[25] Directly constraining the behavior of the AWS allows us to set this objection aside. We also set aside the technical question of whether such a system can be developed and the epistemic question of whether or not it can be verified that a system satisfies these axioms.

Second, we do not assume the AWS is a humanoid construction like the Golem. It is a system that may take on many forms or have many constituent parts, like a drone swarm. In other words, AWS may be closer to Israel's Iron Dome than the Terminator. Third, even an AWS that satisfies all of the Golem Axioms may not be permissible to use in all situations. Other conditions, like consent from the vulnerable community, may need to be met. Finally, the Golem Axioms are in no particular order, as we believe each axiom is necessary and no axiom takes priority over another.

<u>Axiom 1: Fully Autonomous</u>

While Yossele was given tasks to complete by the Maharal, he was able to act independently toward his goals. Similarly, once the system's ends are set, we assume it can act fully autonomously without needing real-time human control and decision-making. While the AWS is expected to communicate with humans, it is not being

actively controlled like an uncrewed vehicle. In Crootof's taxonomy, the AWS is autonomous—not merely automated or semi-autonomous.[26]

Axiom 2: Able to Learn and Adapt to New Situations

Yossele's ability to receive information and adjust plans accordingly was instrumental in protecting Prague's Jews. We assume the system's behavior is flexible enough to learn and adapt to new situations, make observations, and deduce appropriate courses of action in real- time.

Axiom 3: Able to be Given New Instructions

The Maharal regularly redirected Yossele toward where he was needed most. Similarly, as a real-world conflict evolves, the objectives of the AWS may need to change. We assume the AWS can be given new objectives and instructions anytime, including being shut off. We view this axiom as existing in the broader framework of meaningful human control; in particular, delivering new instructions is a form of "human on the loop" control.[27]

Axiom 4: Cannot be Hacked

Yossele only takes commands from the Maharal and his wife. We interpret this as the Golem being unable to be "hacked" and redirected toward inappropriate tasks. Thus, we assume that the AWS cannot be hacked or modified to no longer conform to the Golem Axioms. This axiom eliminates a potential technological risk (not moral) in deploying AWS.[28]

Axiom 5: Free of Bias

Yossele's ability to navigate the world was not hampered by bias and stereotypes—in particular, he did not hold prejudice toward gentiles. Similarly, we

assume that the salient social characteristics of those humans present do not hinder the AWS's capacity to conform to these axioms. In particular, the system's ability to distinguish between friends and foes is not impacted by race, religion, dress, sex, gender, or any other meaningful characteristic. In addition to being an engineering challenge for today's AI systems, algorithmic bias has been used as an argument against the deployment of AWS.[29]

Axiom 6: Reliable Friend vs. Foe Discrimination

Yossele never confused an ordinary gentile for a threat to the Jews of Prague. Thus, we assume the AWS can effectively distinguish between the people it is instructed to protect, active threats, and third parties. This precludes the possibility of the AWS attacking the protected group or other non-threats, like members of an NGO providing aid. Further, we assume that AWS can determine if an actor is a legitimate target, an important distinction noted in the literature.[30]

Axiom 7: Specified Zone of Use

Yossele was not a protector of the Jews of the *world* but a protector of the Jews of *Prague.* Even after Prague was safe, the Maharal did not send him out into the world. We similarly assume that the AWS has a precise, predefined area of operation. This can include different rules of engagement in different sub-areas.

Axiom 8: Lack of Sentience and Other Qualifiers for Moral Worth

Rosenberg explicitly states that Yossele does not have a divine soul. We assume that the AWS is not sentient, has no conscious experience, feels pain or remorse, or experiences psychological trauma. As such, we need not be concerned about the harm to the AWS in a direct moral sense—damaging or destroying the AWS

is not as morally wrong as harming an agent. However, this could still be considered morally wrong if it increases the risk to the vulnerable community.

Axiom 9: Alignment with Moral Principles

The AWS follows a moral code, which entails following international law. The AWS does not use excessive force beyond what the situation demands, desecrate corpses, torture captured combatants, or hurt civilians. Further, the AWS is sensitive to local culture and values. For example, even when tactically effective, when possible, the AWS will avoid combat in and around places of great cultural significance. Maintaining a "defensive posture" is core to these moral principles.[31] Compliance with Axiom 9 requires various stakeholders—representatives of vulnerable communities, the AWS developer, the AWS donor, and the AWS commander—to align their values and objectives. In Rosenberg's narrative, these roles were unified in the Maharal. However, in today's reality, these stakeholders are distinct entities. The success of AWS deployment depends on value alignment among these parties, underscoring the need for collaboration, particularly with those representing the vulnerable community.

In the introduction, we loosely defined DAWS as a distinct category of autonomous weapon systems that only act defensively. For the remainder of the paper, we use "DAWS" to refer to systems that satisfy the Golem Axioms. While it remains unclear if such a system could ever be developed or if we could ever know that a DAWS satisfies these axioms, idealizing DAWS can help clarify the theoretical issues with the system's deployment. Additionally, we believe that an AWS that only partially satisfies the Golem Axioms (or satisfies weakened versions of the Golem Axioms) would still merit the descriptor of DAWS. In particular, our arguments about the

Responsibility to Protect doctrine and humanitarian intervention do not require the full strength of the Golem Axioms.

The Golem Axioms axioms, especially the ninth, are designed to make a DAWS more valuable to a defensive war effort than an offensive one. Suppose both the persecuting group and the vulnerable community have access to a DAWS that satisfies the Golem Axioms. Axiom 9 ensures that both DAWS follow the moral code and maintain defensive postures through the conflict. The vulnerable community's DAWS can actively work defensively to stop the persecution and use lethal means when necessary. However, the persecutor's DAWS can only engage defensively when the vulnerable group oversteps the boundaries of defensive war. It follows that the DAWS is fundamentally more efficacious for the defensive war effort than the offensive war effort. We refer to this difference in efficacy as the Offensive/Defensive Utility Gap (ODUG).

The ODUG does not imply that a DAWS that satisfies the Golem Axioms would provide no value to an unjust war effort. First, such a DAWS could still assist an unjustified defensive position in war (if such a position exists). Second, the DAWS could indirectly assist in an offensive war effort, a possibility we discuss later in the paper. What is important, though, is that the DAWs will not directly contribute to an offensive war effort. If the persecutor has an AWS that aids in persecution, this AWS would necessarily fail to satisfy the Golem Axioms.

Use of Force and Five Cases

When and how much force is justified for a DAWS to use? Axiom 8 specifies that it is pro tanto permissible to destroy a DAWS. Consequently, justifying force through self-defense is suspect. If a DAWS is to use force—especially lethal force—it must be grounded in other-defense.[32] This section explores five cases where a DAWS could be deployed to safeguard a vulnerable community. We use the Golem Axioms to guide our understanding of the system's actions. These cases aim to demonstrate how the force used by a DAWS is proportional to the risk posed not to the DAWS itself but to the vulnerable community.

Humans

In *Humans*, unjust combatants present an imminent threat to civilians or just combatants, such as indiscriminately bombing an urban center or military base that houses just combatants. We take *Humans* as a clear-cut case where a DAWS is justified in using force, including lethal force, to protect civilians and just combatants. If a DAWS is not justified in using force in this case, then it is not clear that it is ever justified in using force.

System

In *System*, unjust combatants are attacking a DAWS in a way that threatens the entire system's integrity. For example, unjust combatants are shooting rockets at the DAWS, and a successful strike will render the DAWS unable to protect civilians and just combatants. In this case, it also seems clear that force, including lethal force, is justified in the interest of protecting innocents. However, without humans in imminent danger, the option of disengagement raises the bar for lethal force. Suppose the combatants attacking the DAWS would pose a threat to innocents only if the DAWS is

destroyed. Further, suppose the DAWS could easily protect itself by taking cover. In this situation, the DAWS should take avoidant actions.

Suppose, however, that there is no way for the DAWS to defend itself without using lethal force. If the DAWS does not defend itself, the unjust combatants will be positioned to harm innocents. In this case, self-defense by the DAWS is required for the continued defense of civilians and just combatants, thus making lethal force a proportional, morally justified response.

Component

In *Component*, unjust combatants are attacking one or more of a DAWS's components, but destroying the component(s) would not hinder the DAWS's ability to protect. The DAWS could sustain this degree of damage without becoming combat-ineffective. Suppose the DAWS has the form of a drone swarm. If an aggressor shoots down an individual drone, this action, while damaging, would not significantly impede the DAWS. Since no human lives are at stake and the DAWS is not at risk of being destroyed, the DAWS is not justified in using force against the unjust combatant merely to protect its parts.

However, if destroying parts of the DAWS diminishes its capacity to protect, it is justified in using proportional force to defend its parts. Proportional force includes non-lethal means of defense, like rubber bullets, bean bag rounds, stun grenades, tasers, and sonic weapons. These non-lethal alternatives are designed to inflict pain or incapacitate unjust combatants and would serve as a way of stopping the attack without resorting to lethal force.

Latent

In *Latent*, unjust combatants are occupying an area near a vulnerable population but are not actively threatening civilians or just combatants. Instead, they are settled near vulnerable humans and can move to harm them at any time. One possible action for the DAWS might be observing the situation until the hostile forces withdraw or begin to act aggressively. However, since a DAWS is a finite resource, it is limited by how much time and how many components can be allocated to monitor potential threats. Adversaries might strategically disperse their forces to stretch the DAWS thin and ultimately prevent it from effectively protecting the vulnerable community. Therefore, the DAWS must find a way to address this situation actively.

In this case, the DAWS would be justified in proactively addressing the potential threat using non-lethal methods. Like *Component*, this case highlights the importance of equipping AWS, broadly construed, with non-lethal capabilities. These capabilities ensure that a DAWS can compel hostile forces to vacate an area.

Vengeance

In *Vengeance*, the vulnerable community eventually defeats the aggressors. Consumed by anger, they start an offensive war against the former aggressors, reversing the roles of victim and victimizer. The offensive campaign's motivation is irrelevant; what matters is that the offensive/defensive dynamic has inverted. Under these circumstances, the DAWS would not become a tool for revenge. While it could maintain a defensive posture by actively monitoring for and confronting any legitimate threats that may arise, Axiom 9 ensures that the DAWS will not directly contribute to

this new offensive war effort. As soon as the offensive/defensive dynamic has inverted, the combat value of the DAWS diminishes due to the ODUG.

Vengeance illustrates how DAWS interacts with the idea of *jus ad bellum*. In the first stage of the war, the DAWS actively supported a just, defensive war. However, when the previously vulnerable group initiates an unjust war of vengeance, the DAWS ceases direct support. Nevertheless, the DAWS can still maintain a defensive posture by being *prepared for* another defensive conflict.

Objections and Responses

The Golem Axioms allow us to set aside some common arguments against AWS, such as unpredictability, bias, and target discrimination. While these arguments are important, setting the usual controversies aside and focusing on a purely defensive system that more closely aligns with moral intuitions and legal conventions opens the door to new objections. In this section, we raise and respond to two potential objections to DAWS regarding the fungibility of military resources and potential misuse.

Military resources are generally fungible because specific resources can be substituted so long as military objectives are still met. Despite training, location, and suitability concerns, soldiers and equipment are generally interchangeable. Consequently, a soldier freed from one role or operation can be immediately directed toward another task. This fungibility has implications for introducing new, more efficient military technology, like DAWS.

Providing a DAWS to a people fighting for self-preservation would likely decrease the need for soldiers patrolling and protecting civilian population centers. The soldiers previously occupying these defensive roles would then be free to occupy

offensive roles. This relative increase in potential offensive resources may increase the likelihood of a just war of self-preservation becoming vengeful and unjust. Further, a DAWS offers this same freeing of resources to the offensive side as well, increasing their capacity for aggression. According to this objection, this dynamic closes the ODUG of DAWS.

Since most military resources are fungible in this way, this is not specifically an argument against DAWS but against any foreign military aid to a group defending its people or sovereignty. For example, providing conventional troops and arms frees resources that can be used offensively. Moreover, this objection applies beyond direct military aid; since money is fungible and military resources can be purchased, aid in cash, food, and humanitarian resources meets the same objection. If fungibility makes providing a DAWS to a group defending its people or sovereignty impermissible, it also renders all direct military, monetary, and humanitarian (e.g., food and medical supplies) aid impermissible. While the fungibility objection reveals a limit of the ODUG, it does not close the gap. Ultimately, this objection proves too much and renders all direct military aid and humanitarian aid impermissible.

A closely related objection is that it may be possible to use DAWS to wage a war of territorial expansion. For example, a state might move civilians into territory captured with conventional troops and then use DAWS to protect these civilians and the newly expanded border. So, while the DAWS is *de jure* defending civilians, it is *de facto* defending an encroaching border.

Like our response to the previous objection, this is not a DAWS-specific problem. A state can use conventional troops to protect civilians in a territory gained

through an illegal war of expansion. However, even if this was DAWS-specific, there are ways to address this concern directly and preemptively. For example, suppose a sponsor like the UN or NATO provides the DAWS. This supranational body would pre-specify an authorized zone of use for the DAWS, as required by Axiom 7. Outside those borders, the DAWS will not operate. If necessary, the sponsor could change the zone of authorized use to stop the encroachment.

Defensive Autonomous Weapon Systems and the Responsibility to Protect Doctrine

Until now, we have been discussing a general use case for DAWS to protect vulnerable populations and exploring some normative and theoretical questions along the way. In this section, we demonstrate how DAWS can help resolve three practical challenges related to the Responsibility to Protect (R2P) doctrine and some normative challenges to humanitarian intervention.

R2P arose in response to the international community's failure to prevent and stop genocide, war crimes, and crimes against humanity in the twentieth century.[33] The doctrine was first introduced in 2001 by the International Commission on Intervention and State Sovereignty (ICISS) after recognizing the need for a new approach to respond to the threat of mass atrocities and the lack of an international responsibility to protect vulnerable populations. In 2005, the United Nations officially endorsed the R2P doctrine, and it has since been used as a tool to prevent, halt, and address mass atrocities such as genocide and ethnic cleansing. The doctrine provides a framework for the international community to take collective action to prevent and respond to

gross human rights abuses and to protect vulnerable populations when national authorities are unable or unwilling to do so.

R2P is based on the principles of sovereignty, non-intervention, and the state's responsibility to protect its citizens.[34] At the same time, it recognizes the international community's collective responsibility to protect populations from mass atrocities. As a normative framework, R2P has gained widespread acceptance within the international community.[35]

Although R2P has a sound theoretical foundation and broad political support, its practical results have been mixed. The 2011 intervention in Libya demonstrated the successful implementation of R2P. The UN Security Council authorized a group of nations, led by the United States, to step in and safeguard civilians from Colonel Muammar Gaddafi's oppressive regime. This intervention aimed to prevent the widespread violations of human rights and the threat of imminent violence that Gaddafi's regime was responsible for.[36] This successful intervention ultimately led to Gaddafi's overthrow and the restoration of stability in Libya, highlighting the international community's commitment to uphold the protection of civilians and prevent the escalation of violence.[37]

However, R2P was largely unsuccessful in preventing mass atrocities during the Syrian Civil War.[38] Although the UN Security Council passed multiple resolutions to protect the civilian population, disagreement among major powers, specifically the United States and Russia, prevented consensus on a unified approach.[39] The

international community's failure to act had catastrophic repercussions: hundreds of thousands have been killed, and millions more have been displaced.[40]

R2P faces three significant hurdles. First, due to the associated costs, domestic political considerations may prevent a well-resourced state from participating in a humanitarian intervention.[41] Additionally, providing weapons and training to states or local resistance forces carries significant risk.[42] Lastly, deploying traditional troops can result in negative externalities for populations in the affected areas.[43]

We suggest that using DAWS in place of traditional military intervention or arming and training vulnerable populations offers a more ethical, effective, and economical approach to protecting these communities from mass atrocities. Specifically, deploying DAWS may result in fewer casualties, be less risky than providing arms and training to local fighters, and be more cost-efficient. Additionally, an autonomous system may impose fewer negative externalities on local populations than traditional troops.

Less costly than traditional military intervention

In the future, deploying DAWS to protect vulnerable communities may be less costly—in so-called blood and treasure—than traditional military interventions.[44] Substituting conventional troops with DAWS will likely reduce the number of casualties for sponsors, the amount of political capital that leaders must expend domestically, and the total monetary cost.[45] These costs are often limiting factors that can reduce the likelihood of a sponsor taking action to protect a vulnerable community. Lowering or eliminating them altogether can increase the potential for quick and decisive action.

Reducing sponsor casualties is an important concern for both normative and practical reasons. Normatively, it is difficult to determine the extent to which soldiers have a moral obligation to protect the lives of citizens in foreign countries.[46] Practically, the possibility of casualties can deter global leaders from (a) intervening in the first place and (b) sustaining troop presence long enough to prevent the open-ended chaos that political scientists and military planners fear.

Global leaders may lack the political will to conduct a timely and effective humanitarian intervention. The Clinton Administration's failure to intervene in Rwanda in 1994 is a poignant example of the lack of political will to intervene despite escalating violence, in part due to the political backlash of the casualties in Somalia two years prior.[47] If using DAWS reduces human casualties on the battlefield, this political hurdle may be reduced or possibly eliminated.

Deploying human warfighters is expensive; the estimated costs of the United States' wars in Iraq and Afghanistan range from $4 to $6 trillion.[48] Citizens in sponsor states do not bear unlimited financial responsibility to citizens in other states. Using DAWS may reduce the cost of protecting human rights, like how drones have reduced the cost of bombing and reconnaissance missions.[49] All else equal, a less expensive military intervention is more politically acceptable and financially just than a more costly operation.

Indeed, a common argument against traditionally conceived AWS is that they lower the threshold for war.[50] If an AWS is more easily deployable than traditional military forces, states may be more likely to engage in war, or so the argument goes. However, there are two sides to this coin; lowering the threshold for war also lowers

the threshold for humanitarian intervention. When there is little political will for a conventional troop deployment, the choice may be between deploying DAWS or allowing the mass atrocity to continue. For instance, if the international community in the future faces a situation like the 1994 Rwandan genocide, and there is no political will to support military intervention, but there is political will to deploy DAWS, then deploying DAWS may be the best practical humanitarian response.

Less risky than providing traditional arms and training

Since traditional arms are prone to misuse and are hard to reclaim, arming local groups is a risky endeavor.[51] For example, the United States' arming of the Taliban in the 1980s eventually resulted in arms being used against American soldiers in the early 2000s.[52] Moreover, hostile groups may even seize weapons or equipment (e.g., Iraq in the early 2010s and Afghanistan in 2021), leading to a loss of control over these resources. Moreover, while so-called dumb weapons, like rifles, can be effective in the hands of trained soldiers, an untrained individual may not be equipped to take full advantage of their potential. Providing arms to vulnerable communities without proper training may not effectively achieve the desired objectives. Providing adequate training demands significant time, resources, and effort, which might not be practical in some situations. Moreover, training local fighters is risky—trained warfighters may use that training in ways the trainer state never intended.

DAWS offers a potential solution to the issues above, as it can reduce the risk of misuse and allow the sponsor to maintain greater control over the use of force. DAWS can be designed for external control, allowing for the safe decommissioning of the system once the mission is complete. This would prevent the vulnerable community

from engaging in hostile actions beyond what is necessary for their protection. Moreover, external controls would allow sponsors to reclaim the systems (e.g., the system drives or flies home), thus reducing the likelihood that a bad actor would gain control over and potentially repurpose the system.

Fewer negative externalities on local populations

Traditional troop deployments into a local community are highly visible due to the troops' physical presence, identifying uniforms, foreign language, and weapons. Traditional soldiers are a constant reminder of foreign presence and may generate animosity from the local population and even encourage increased violence.[53] However, unlike human soldiers, DAWS might operate with relatively low visibility. For example, a hypothetical defensive drone swarm could protect a street market from high above without creating a noticeable physical presence on the ground.

Additionally, traditional troops can act in ways that harm the local population. Perhaps most obviously, even when fighting in a just war, rogue human soldiers might commit war crimes such as rape, murder, torture, wanton destruction, and corpse desecration. Even assuming that the majority of soldiers refrain from committing illegal actions in just wars, it is still unrealistic to expect that war crimes by human soldiers can be eliminated. Our suggestion here is not that a DAWS could never commit a heinous act but rather that such an act would be less likely to occur than under a traditional troop deployment. Moreover, certain war crimes, like rape, would be vanishingly unlikely.

Beyond direct criminal behavior, the influx of foreign military personnel, primarily young men, into an area can result in negative social externalities for the local

population, especially women. For example, deploying traditional troops may create or expand black markets for sex and drugs. The taking of war brides is a common practice in conflict zones. It has been documented extensively in recent conflicts, such as the Iraq War, where American soldiers took an estimated 2,400 war brides.[54] Additionally, the presence of a high number of young men has been linked to the proliferation of sex work economies, most notably in Southeast Asia. During the Vietnam War, prostitution levels rose substantially in South Vietnam, leading to a range of adverse social issues, such as the mistreatment of Amerasian children born to sex workers.[55] Because it is not human, a DAWS would not engage in these uniquely human vices.

There are, of course, some cases where DAWS cannot prevent serious human rights violations. For example, it is unrealistic to expect that DAWS could prevent human rights violations when its deployment would be seen as a direct violation of the sovereignty of a military superpower. However, given that the current system is ineffective at preventing such crimes, this objection does not hold. Our argument is not that DAWS is a one-size-fits-all solution to a complex problem such as protecting human rights. Instead, traditional tools of diplomatic engagement, economic sanctions, and other measures remain necessary to address human rights issues that arise. We want to suggest that DAWS could be an essential tool in the humanitarian toolkit in the future.

Conclusion

The Maharal of Prague created his Golem with a specific mission: “[G]uard the Jews from all harm and from all the ills and troubles they suffer at the hands of their enemies and oppressors.”[56] We envision a future where DAWS will be given similar directives to protect vulnerable communities from persecution. These systems may one day enable more effective humanitarian interventions and realize the lofty normative ideals embodied in the Responsibility to Protect doctrine. If effectively deployed, a DAWS may even serve as a deterrent to would-be oppressors, much like how word of Yossele deterred antisemites outside of Prague.

Our discussion may serve as a moral justification for AWS development in the interest of developing DAWS, but we have not explored this argument here. Once developed, will AWS play the role of Terminator or Golem? Ultimately, the case for AWS development must carefully weigh the potential benefits, such as DAWS protecting vulnerable groups, against the risks, like misuse leading to harm. We view this paper as a contribution in favor of DAWS development.

After decommissioning Yossele, the Maharal placed his body in the synagogue's attic, anticipating that there might come a time when Prague's Jews would need their protector once more. In today's world, numerous vulnerable communities desperately need protection. We view the advancements in artificial intelligence and robotics as the loam and clay for a modern Golem.

---

[1] See Horowitz, "The Ethics & Morality of Robotic Warfare," Liao, "A Short Introduction," Asaro, "Autonomous Weapons," and Wendell and Valor, "Moral Machines," among others.

[2] Jacoby, “The Golem in Jewish Literature.”

[3] Schäfer, “The Magic of the Golem.”

[4] Leiman, “The Adventure of the Maharal.”; Leiman, “Did a Disciple of The Maharal Create a Golem?”

[5] Jacoby, “The Golem in Jewish Literature.”

[6] Schäfer, “The Magic of the Golem”

[7] Scholem, “The Idea of the Golem.”

[8] Leiman, “The Adventure of the Maharal.”; Dekel, Eden, and David Gantt Gurley. “How the Golem Came to Prague.”

[9] Scholem, “The Idea of the Golem.”

[10] Dekel and Gurley, “How the Golem Came to Prague.”

[11] Scholem, “The Idea of the Golem.”

[12] Dekel and Gurley, “How the Golem Came to Prague.”

[13] Rosenberg and Leviant, *The Golem.*

[14] Ibid.

[15] Jacoby, “The Golem in Jewish Literature.”

[16] Glinert, “Golem!”

[17] Contrada, “Golem and Robot.”

[18] For humanitarian organizations and NGOs, see *Human Rights Watch* (Docherty, “Losing Humanity.”) and *The Campaign to Stop Killer Robots* (Abi Assi, “Stop Killer Robots.”; PAX, "Save Your University."). For examples from philosophical literature, see Sparrow, “Killer Robots.”; Young, “Indignity of Killer Robots.”; Krishnan, *Killer Robots.*

[19] “Blood libel" is not merely a part of the fictional setting but rather a historically common and particularly harmful accusation against Jews. See Gottheil, Strack, and Jacobs, “Blood Accusation.”

[20] While Rosenberg’s home of Warsaw was relatively safe in the years leading up to the publication of the novel, there was notable nearby antisemitic violence. For example, an estimated 400-800 Jews died in a single 1905 Pogrom in Odessa (Weinberg, “The Pogrom.”) and Leopold Hilsner faced public trial in Austria-Hungary after a false accusation of ritual murder (Gotthard, “Polna Affair.”)

[21] The linguistic similarity between circuits as laps and circuits as electrical wiring should not be ignored. Some Kabbalistic traditions believe linguistic similarity reveals similar underlying structures (Idel, “Reification of Language.”)

[22] Rosenberg and Leviant, *The Golem.* 36

[23] See Williams, *Moral Luck;* Arpaly, “Moral Worth.”

[24] Asimov, *I Robot;* Arkin, Ulam, and Duncan. “An Ethical Governor.”

[25] Sparrow, “Killer Robots.”; Asaro, “Autonomous Weapons.”

[26] Crootof, “Autonomous Weapon Systems.”

[27] For a recent overview of meaningful human control, see Amoroso and Tamburrini, “Meaningful Human Control.”

[28] Asaro, “Autonomous Weapons.”

[29] Ramsay-Jones, “Racism and Fully Autonomous Weapons.”

[30] Sparrow, “Robots and Respect.”

[31] The line between offensive and defensive action is often clear in the abstract but messy in practice. We suggest that supranational institutions such as the United Nations could determine these distinctions in real-world scenarios.

[32] For an overview of other-defense, see: Frowe, *Ethics of War and Peace*. 24-26.

[33] Global Centre for the Responsibility to Protect. "What is R2P?"

[34] Ibid.

[35] Evans, "2020 Annual Lecture."

[36] United Nations. "Security Council Approves 'No-Fly Zone.'"

[37] Adams, "Libya and the Responsibility to Protect."

[38] Williams, Worboys, and Ulbrick, "Preventing Mass Atrocity Crimes."

[39] Ibid.

[40] Al-Oraibi, "Syria's Decade of Conflict."

[41] Baldauf, "Why the US didn't intervene in the Rwandan genocide."

[42] Pattison, "The Ethics of Arming Rebels."

[43] Moon, *Sex Among Allies;* Baker, *American Soldiers Overseas.*

[44] This argument has been made to support both remote warfare and traditionally conceived AWS. See Sparrow, "Robots and Respect;" Frowe, *Ethics of War and Peace*, Chapter 11; and Riesen, "The Moral Case."

[45] We assume the long-run cost of deploying a DAWS will be less than recruiting, training, equipping, sustaining, and caring for human soldiers during and after combat deployments. We acknowledge that this depends on myriad factors (e.g., production, raw materials cost, etc.).

[46] Walzer, *Just and Unjust Wars.*

[47] Baldauf, "Why the US didn't intervene in the Rwandan genocide."

[48] Bilmes, "The Cost of the Iraq War."

[49] McLean, "Drones are cheap."

[50] Blanchard and Taddeo, “Autonomous weapon systems and *jus ad bellum*.”

[51] Pattison, “The Ethics of Arming Rebels.”

[52] Silverstein, "Who's Got the Stingers?"

[53] Pape, Robert A. “The Strategic Logic of Suicide Terrorism.”

[54] Dickey, and Ramirez. "Battlefield Romances in Iraq."

[55] Yarborough, *Surviving Twice.*

[56] Rosenberg and Leviant, *The Golem.* 36