An International Journal in Philosophy, Religion, Politics, and the Arts

Existenz

Volume 17, No 1, Spring 2022 ISSN 1932-1066

LAWS and the Law: Rules as Impediment to Lethal Autonomy

Kevin Schieman
United States Military Academy, West Point
kevin.schieman@westpoint.edu

Abstract: The ability to follow the laws of war is arguably a necessary, if not sufficient condition for morally acceptable lethal autonomous weapons use. Yet following laws and other types of abstract rules is far more demanding than one may realize. This type of rule-following requires an agent to overcome three main difficulties: the heterogeneity of rules, semantic variation, and the demands of global reasoning. Taken together, these demands set an ambitious agenda for research focused on building machines capable of satisfying even a modest standard of legal compliance.

Keywords: Military ethics; machine ethics; lethal autonomous weapons; *jus in bello*; artificial intelligence; law; moral agency; machine agency.

Recent technological advances have led to a surge of optimism about the near-term prospects for many applications of artificial intelligence and machine learning.[1] Especially significant have been breakthroughs in computer image recognition and natural language processing, hard problems with substantial practical utility that once seemed prohibitively difficult to resolve computationally. This progress has given rise to speculation that humans have reached an inflection point in artificial intelligence research where the technology has advanced to such a level that intelligent machines will begin to assume weightier roles in our lives. Perhaps nowhere has this been more obvious than with the public release of the artificial intelligence chatbot ChatGPT, a large language model with a somewhat general ability to complete a range of tasks in response to natural language prompts. It is hardly an exaggeration to say that the technology has caused a near panic in higher education and even led some to lament the impending death of the scholarly essay. Similarly important are advances in other domains of research such as self-driving cars, which appear to be making measurable progress in their capacity to cope with the irregularity and unpredictability of driving on public roads. It is in view of this progress that I want to revisit an important argument from the literature on Lethal Autonomous Weapon Systems (LAWS); a LAWS, according to the U.S. Department of Defense, is a system that

> once activated, can select and engage targets without further intervention by an operator.[2]

[1] Disclaimer: The views expressed in this paper are those of the author and do not reflect the official policy or position of the United States Military Academy, the Department of the Army, the Department of Defense, or the United States Government.

[2] Kathleen Hicks, *Autonomy in Weapon Systems (DoD Directive 3000.09)*, Washington, DC: 25 January 2023, p. 21, https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf

Specifically, I shall consider the argument that armies morally ought to adopt LAWS, so long as those systems are capable of following the laws of war. One way of arriving at this conclusion is to contend that wars fought by machines whose behavior is compliant with the laws of war would be morally better than the *status quo ante machina*, in which wars are fought by human soldiers that often kill indiscriminately or otherwise commit battlefield atrocities.[3] Whether or not such a state of affairs would in fact be better is ultimately an empirical question but notice that the argument ascribes great moral importance to following the laws of war. That is to say, the claim that humans would be morally better off with legally compliant LAWS depends upon the belief that the laws of war have a certain moral content and that in following those laws, soldiers act morally, or at the very least, they avoid acting immorally. This suggests that the ability to follow the law is necessary for soldiering well and by extension, would likewise be necessary for any morally acceptable LAWS. This is an important moral claim, and it presents a number of complicated philosophical questions. For example, it seems obvious that one can follow the law, yet still act immorally. If this is the case, then soldiers ought to aspire morally to more than legal compliance, even if they sometimes fail to achieve the legal standard. Rather than address such questions here, I want to proceed on the assumption that the ability to follow the law is prerequisite to the possibility of there being ethical LAWS. Granting that assumption, I want to consider the practical difficulties inherent to building such a machine.

An analysis of the demands imposed on moral agents by following the law shows that the laws of armed conflict and other relevant legal guidance are variously abstract and hierarchical, qualities that make legal rule-following especially difficult. As a result, following the laws of war requires an agent to overcome conceptual challenges posed by the heterogeneity of rules, semantic variation, and the demands of global reasoning, to say nothing of the difficulty of grasping the concept of law itself. It turns out that following abstract rules is far more difficult than it is generally given credit for being, most likely because humans have failed to properly appreciate their own impressive abilities to understand and to follow rules. As Marvin Minsky once remarked, "in general, we're least aware of what our minds do best."[4] It may be that even if states ought to build law-abiding machine-weapons, realizing that goal might turn out to be quite technically demanding. None of this is to suggest that these problems are insuperable from the perspective of artificial intelligence research—ultimately, that is also an empirical question. Instead, the point is to emphasize the ambitiousness of many of the goals in artificial intelligence research and to stress the importance of remaining clear-eyed in pursuit of those goals, especially when the real-world stakes are as consequential as they are with robotic weapons.

The *Status Quo Ante Machina*

In his 2009 book, *Governing Lethal Behavior in Autonomous Robots*, Ronald Arkin makes a case for developing lethal autonomous weapons on the grounds that they might follow the laws of armed conflict more consistently than human soldiers do. Since soldiers frequently violate these laws, Arkin reasons that the standard for robotic weapons need not be perfection, but rather, mere improvement from the *status quo ante machina*. He cites an official United States Army mental health survey of soldiers returning from the war in Iraq that captures troubling attitudes toward Iraqi civilians and non-combatants, including permissive attitudes toward abuse and torture (*GLB* 31-2). Of course, such attitudes are hardly isolated to any particular army or war. He concludes that the possibility of building machines capable of satisfying the modest, but morally important standard of legal compliance is a sufficient reason to invest into this technology. Frankly, if one were looking for a good moral reason to adopt lethal autonomous weapons, it is hard to imagine a better one than their potential for reducing avoidable harms in war.

Given the imperfect standard set by human soldiers, Arkin reasons that it is at least possible that humans might one day—and maybe even someday soon—be able to build machines that do at least as well as they themselves do, and maybe a great deal better. In contrast with human soldiers, he explains, robotic weapons would not suffer potentially error-inducing emotions such as anger, fear, or frustration; they could be programmed to exercise more conservative

[3] Ronald C. Arkin, *Governing Lethal Behavior in Autonomous Robots*, Boca Raton, FL: Chapman & Hall/CRC 2009, p. 31. [Henceforth cited as *GLB*]

[4] Marvin Minsky, *The Society of Mind*, New York, NY: Simon & Schuster 1986, p. 29.

judgments about collateral harm; they could be designed to incorporate better sensors and more real-time information processing than human soldiers; and they would not be subject to the same cognitive biases that sometimes undermine human judgment (*GLB* 29–30). For such reasons, Arkin is not alone in advocating for the potential moral benefits of LAWS, and similar arguments have been advanced by others, including recent work by Don Howard.[5] Whether these differences would indeed lead to greater compliance with the laws of armed conflict is hard to say, however, the dangers in presuming too much of the technology seem obvious. As much as one should care deeply about the ways that soldiers fail morally, it would be a mistake to ignore the many amazing things that they are capable of doing in virtue of their being human. Arkin is right to point out that human soldiers sometimes fail to follow the laws of war, but with rare exceptions, they fail to do so despite having the ability to understand what the law requires of them. Even when they fail to follow the laws of war, soldiers achieve something that machines thus far have been generally unable to accomplish: In understanding the ways that the law obligates them and constrains their actions, soldiers achieve something that is morally important, even if they fail to make good use of that achievement. Machines may eventually be capable of the same kind of understanding, but quoting my colleague, Richard Schoonhoven, at this point machines "aren't even stupid."[6]

Making Sense of the Law

The most sensible starting point for analyzing the demands that following the law makes on agents is the concept of law itself. The benefit of starting there, I hope, is to make clear that even the concept of law is not simple—the law consists in various types of rules and rule-like contents, each of which imposes different demands on its subjects. In this section, I briefly describe three categories of legal contents, before turning to the ways in which the heterogeneity of laws makes legal rule-following especially difficult. In his analysis of law, Paul Boghossian identifies three distinct types of content that compose law: rules, imperatives, and normative propositions.[7] On most philosophical analyses of law these three contents compose the law in some measure, but there is significant disagreement regarding the relative contributions of each one. John Austin, for example, argued that laws are the commands (or imperatives) of a sovereign ruler backed by force of sanction, although he recognized that some commands take the more general form of rules.[8] Alternatively, H. L. A. Hart argued that laws are social rules, drawing a distinction between those rules that guide conduct and those that allow the creation or modification of new laws.[9] In stark contrast with both Austin and Hart, Ronald Dworkin argues that in addition to rules and normative propositions, the law consists not only in rules, but in moral principles and policies, meaning that there are aspects of any legal system that cannot be explained strictly in terms of social facts.[10] The point is that there is significant disagreement regarding the concept of law and the role that rules, imperatives, and normative propositions play in constituting the law as such. Importantly, each of these perspectives suggests that there are distinct difficulties for the legal subject charged with, knowing, understanding, and applying the law. Hence, it is important to address the similarities and differences between these types of legal contents.

Boghossian argues that imperatival contents, sometimes called commands or instructions, specify both a condition and an action that is to be performed whenever that condition obtains (*ER* 474). Imperatives tend to be highly specific in their formulation of both the applicable condition and the requisite behavior.

[5] Don A. Howard, “In Defense of (Virtuous) Autonomous Systems,” *Dakota Digital Review*, 21 February 2023, https://dda.ndus.edu/ddreview/in-defense-of-virtuous-autonomous-systems/.

[6] The quote is from a presentation that Schoonhoven gave at the International Society of Military Ethics Conference in 2022. He was riffing on a well-known quote by theoretical physicist Wolfgang Pauli, who once quipped about the work of a young scholar that “It is not even wrong.”

[7] Paul A. Boghossian, "Epistemic Rules," *Journal of Philosophy* 105/9 (September 2008), 472–500. [Henceforth cited as *ER*]

[8] John Austin, *The Province of Jurisprudence Determined*, ed. Wilfrid E. Rumble, New York, NY: Cambridge University Press 1995, p. 275.

[9] H. L. A. Hart, *The Concept of Law*, Oxford, UK: Clarendon Press 1994, pp. 59-60, 94.

[10] Ronald M. Dworkin, "The Model of Rules" *The University of Chicago Law Review* 35/1 (Autumn 1967), 14-46, here p. 14.

Consider the following two imperatives, TURING and CHORES:

> TURING: If the value is 1, write 0 and move the tape right.
> CHORES: If the dishwasher is empty, rinse your dish and place it into the dishwasher.

Both imperatives specify a condition and an action, but the authority of either instruction depends principally on the authority of whomever or whatever is issuing the command. That is to say that whether I ought to obey an imperative depends in significant respects on whether the issuer stands in an appropriate authority relationship to me. In the first case adduced above, the authority is that of a Turing table or a program of some sort. In the second case, it is something like house rules; as my mother used to be fond of saying, "my house, my rules." For CHORES, disobedience may be costly, but for TURING it is not even a possibility.

Normative propositions also specify conditions and action guidance, although the authority of the proposition is grounded more broadly in the system of which it is a part. Still, normative propositions are assertions that either grant a permission or levy a requirement. For example, Boghossian discusses the following two normative propositions (*ER* 474-5):

> OPEN: At the beginning of the game, white must make the first move.
> CASTLE: If the board configuration is C, you may castle (make a specialized move involving both the king and one of the rooks).

Obviously, both OPEN and CASTLE are rules of chess, and as such, their normative authority is conditional. If I intend to play chess at all, then I had better abide by OPEN and, if I intend to play well, I had better know when to exercise the permission afforded by CASTLE. John Rawls describes these types of rules as practice rules whereby "practice" refers to

> any form of activity specified by a system of rules which defines offices, roles, moves, penalties, defenses, and so on, and which gives the activity its structure.[11]

Practice rules are those rules that constitute the practice in a literal sense. If, while purporting to play chess, I move a pawn from square to square, jumping multiple pieces within a single turn, it turns out that I am not really playing chess at all (maybe I mean to be playing checkers). All of this is to say, normative propositions are authoritative only insofar as one intends to participate in activities that are in some sense defined by those propositions. Applied to the law, there are interesting disagreements regarding the way in which the law has normative authority over its subjects. Whereas Austin and Hart grounded the normativity of law in force and social practice respectively, for Dworkin, the force of law lies in the moral justifiability of coercive government action—a law has normative force, just so long as the government would be morally justified in using force to enforce it.[12]

It should be immediately apparent that while there is reason to think that imperatives and normative propositions, and maybe even rules, differ in significant ways, these categories do not separate cleanly from one another. My own discussion of normative propositions has already lapsed into talk of rules and something similar can be said for some imperatives. After all, CHORES is exactly the type of imperative that I would characterize as a rule, and I certainly understood them in that way as a child. Boghossian writes:

> in looking at the literature on rules one is struck by two related observations: one is that different notions are often conflated; the other is that it is often hard to see when a dispute is merely *verbal* and when it is substantive.[13]

Regardless of whether the law consists entirely in rules or some combination of rules and other rule-like contents, following the law is not a singular activity that imposes a singular set of demands on an agent. At the risk of putting too fine a point on the matter, the most basic challenge in following the law is the concept of law itself.

A reasonable response at this point is to acknowledge that a lack of philosophical resolution need not always trouble one in practice. For example, despite widespread philosophical disagreement in normative ethics, one can find a surprising amount of agreement about ethics in practice, a point made compellingly by Alasdair MacIntyre some years ago. As it is, most soldiers manage to follow the law

[11] John Rawls, "Two Concepts of Rules," *The Philosophical Review* 64/1 (January 1955), 3-32, here p. 3.

[12] Ronald Dworkin, *Law's Empire*, Cambridge, MA: Belknap Press 1986, p. 93.

[13] Paul A. Boghossian, "Rules, Norms, and Principles," in *Problems of Normativity, Rules and Rule-Following*, eds. Michał Araszkiewicz, Paweł Banaś, Tomasz Gizbert-Studnicki, Krzysztof Płeszka, Cham, CH: Springer Verlag 2015, pp. 3–12, here p. 3.

without cataloguing the subtle differences between imperatives and normative propositions. Frankly, many lawyers most likely practice law without ever wading too deeply into philosophical analyses of law. The difficulty for building law abiding machine weapons is that many of the distinctions highlighted in the philosophy of law are somewhat intuitive to most people but may be quite difficult to formalize in a machine architecture. As a child, I had little difficulty recognizing that the threat of punishment implicit in deciding to ignore my parents' orders—formalized, for example, by the command CHORE—made them very different from the normative propositions governing a game that I had freely chosen to play with friends. The challenge is not in understanding these differences, so much as it is in representing them in a computer model that is adequate to realize legally compliant behaviors in the world. In a 1998 interview Daniel Dennett remarks:

> There's no place for impressionism in creating a computer model or an algorithm…Computers force you to get clear about things that it's important to get clear about.[14]

Perhaps surprisingly, making clear the conceptual distinctions of moral and legal philosophy may be far more important to building ethical LAWS than they are to human soldiers, who can intuitively grasp normative differences although they might struggle to formulate them.

The Heterogeneity Problem

Whether the law consists exhaustively in rule-following, or whether it consists in following rules and other rule-like contents, the variety of legal contents makes law-following more difficult. In the previous section I suggested that one potentially important difference is that laws vary in terms of the source of their normative authority. In this section, I want to emphasize two other important ways that laws differ—as rules of one type or another—that are especially relevant for present concerns. First, laws vary significantly in the generality or specificity of both the circumstances in which they apply, and in the nature of their action-guiding content. Some rules prescribe a very specific act in response to a very specific condition or circumstance. Alasdair MacIntyre describes these rules as maximally concrete such that "there is no such thing as understanding the rule and not knowing how it is to be applied."[15] Other rules are abstract in that they might fully determine some set of more concrete rules.[16] For example, a relatively abstract rule like, "You should always treat people with respect," fully determines any number of rules about how a person ought to treat others in a particular case. Second, rules vary in the extent to which they are either prescriptive or proscriptive; for instance, rules such as CASTLE make allowances for acting in certain ways, while rules such as OPEN levy requirements. Still other rules prohibit or discourage behaviors, although I have not offered an example of either case. It should be enough to say that differences in the ways that a rule guides action, coupled with differences in type and generality, present several distinct challenges for rule-following.

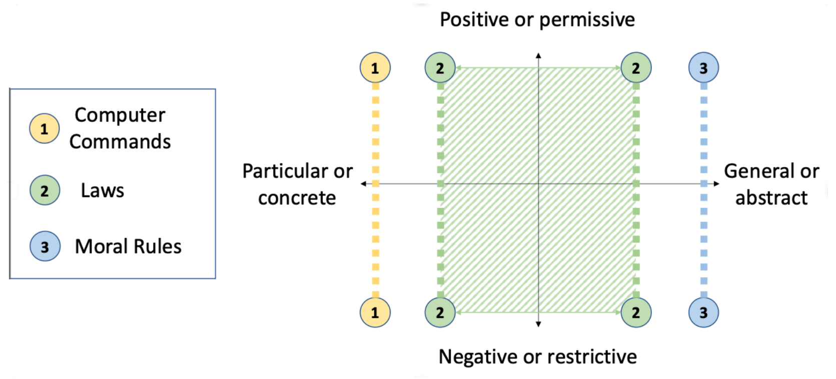

Figure 1: Rules vary according to the permissiveness or restrictiveness of their action guidance (y-axis) and according to the generality of the guidance of the rule (x-axis). Laws tend to occupy an intermediate space in terms of their generality, but admit of both permissive and restrictive cases.

The best approach to demonstrating these challenges may be through direct comparison of different kinds of rules (see Figure 1). Consider, for example, the contrast between computer commands and moral rules. Computer commands, like those that

[14] Harvey Blume, "A Conversation with Daniel Dennett," *The Atlantic Online*, 9 December 1998, https://www.theatlantic.com/past/docs/unbound/digicult/dennett.htm.

[15] Alasdair MacIntyre, "Does Applied Ethics Rest on a Mistake?," *The Monist* 67/4 (October 1984), 498–513, p. 503.

[16] David A. Vogelsang and Mark D'Esposito, Is There Evidence for a Rostral-Caudal Gradient in Fronto-Striatal Loops and What Role Does Dopamine Play?," *Frontiers in Neuroscience* 12/242 (April 2018), 1-11. p. 2.

constitute the operating system of the computer that I am typing this essay on, are imperative contents. These rules tend toward being maximally concrete, and range in their prescriptions from the permissive to the restrictive. On the one hand, they are quite simple, yet their ability to guide even modestly complicated behavior depends upon prohibitively lengthy strings of commands—it is not uncommon for a computer operating system to include on the order of 100 million lines of code. On the other hand, moral rules—or, for Dworkin, principles—can be incredibly abstract, determining action in countless situations, contexts, or practices. For example, a general moral prohibition against harming might determine right action in contexts ranging from war to participation in sports to childhood interactions with siblings. Like computer commands, moral rules may be both restrictive and permissive; Immanuel Kant, for example, thought that humans have an absolute or perfect duty to refrain from lying to others, but a permissive or imperfect duty to cultivate whatever natural talents that they may possess. Many laws seem to lie intermediate with respect to these two cases, benefitting neither from the simplicity of individual computer commands, nor the relative sparsity of moral rules. The reality is that legal contents range from the maximally concrete to the very abstract, whether one characterizes them as rules or as something else, such as imperatives or normative propositions. As such, following the law means dealing with a great deal of variously abstract rules, some of which are quite restrictive, and others of which are quite permissive, but many of which bear on one another in understanding the law *in toto*.

The Semantic Problem

Another problem posed by legal rule-following results from the fact that the laws are written in natural languages and natural languages are notoriously irregular and imprecise. For one thing, rules often admit of an objectionable degree of vagueness, which generally, a rule-follower needs to resolve. This problem seems to become more pronounced as rules become more general because such rules tend to be more dependent upon words with rich semantic content. For example, the words "proportionate," "necessary," "reasonable," and "suffering" each play an important role in law. In the discussion section of a research report conducted by Igor Grossman, et al. the authors agree with findings from a 1993 study by Eldar Shafir *et al.* that colloquial use of the term reasonable "concerns a pragmatic focus on social norms and context specificity in the process of judgment."[17] Grossman, *et al.* note that this is distinct from colloquial judgments about rationality, which appear to be primarily focused on individual preferences and attributes. To the extent that social norms vary across communities, this suggests a real difficulty for rule-following in that the meaning of a rule may similarly differ across communities. Granted, Grossman, et al. focus specifically on the folk use of the word "reasonable" because the term plays an especially significant role in the law, but it would be surprising if social norms did not play a significant role in common understanding of other legally significant terms (for instance, "proportionate," "necessary," and "suffering"). Even limited to the word "reasonable" though, the problem is significant.

Consider, for instance, Article 57 (4) to Additional Protocol I to the Geneva Conventions states:

> In the conduct of military operations at sea or in the air, each Party to the conflict shall, in conformity with its rights and duties under the rules of international law applicable in armed conflict, take all reasonable precautions to avoid losses of civilian lives and damage to civilian objects.[18]

The use of a reasonableness standard in this law is not just problematic on account of some vexing terminological ambiguity; it is objectionable because the meaningfulness of the law depends upon the meaning of the word "reasonable," which is difficult to fix. A high standard of reasonableness directs a highly restrictive rule, while a low standard directs the exact opposite. The entirety of prescriptive work done by the law depends upon a single word and the meaning of that word is incredibly difficult to fix, even for proficient language-users. It is not an exaggeration to say that the rule depends as much on a deep understanding of one specific word's usage within a given linguistic community as it does on the logical relationships between terms in the rule itself.

[17] Igor Grossmann, Richard P. Eibach, Jacklyn Koyama, and Qaisar B. Sahi, "Folk Standards of Sound Judgment: Rationality Versus Reasonableness," *Science Advances* 6/2 (8 January 2020), 1-14, p. 6.

[18] Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts (Protocol I), of 8 June 1977, Art. 57/4, p. 270, https://ihl-databases.icrc.org/assets/treaties/470-AP-I-EN.pdf.

A further difficulty introduced through natural languages is that of conversational implicature, among other discursive norms. Granted, this problem is far worse in conversation than in law, but that does not mean the problem is insignificant, even in that context. For example, when I tell my son, "Don't spit on people," I actually mean something closer to "under no circumstance is it permissible to spit on people." It is only through the practical experience of parenting two small children that I have come to learn that that particular implicature is entirely lost on some members of my own linguistic community. A more interesting example, potentially, is when I tell him "Don't hit people." I suppose that what I probably mean is more like "never hit a person unless they are physically harming you." I think this example is fascinating because it turns out that rules on justifiable harming are rather nuanced. In fact, I find myself struggling to specify precisely the circumstances under which harming others is appropriate. Of course, I should expect my kids to learn over time to account for considerations of preemption, proportionality, and necessity in developing a finer understanding of the rules against hitting others. Still, it is difficult to specify all the morally relevant considerations bearing on the justifiable use of violence, and I take it that these are relatively simple rules. Maybe I could just say, "don't hit others unless it is strictly necessary that you do so." But notice there that the phrase "strictly necessary" has introduced the very sort of ambiguity and subjectivity that one should hope to avoid with children and robotic weapons. I have argued elsewhere that children, especially teenagers, offer a useful parallel in guiding effective lethal autonomous weapons policy, a point which takes on an even greater significance in recognizing the linguistic demands of following the law.[19]

The Global Reasoning Problem

Thus far, the problems that I have described are features of individual laws and legal terms that make it difficult to understand how the law obligates people. However, rules also present a problem collectively that I refer to as the global reasoning problem. Whatever challenges one might find in applying a single rule, rules are rarely, if ever, applied individually. Instead, following the law is usually a practice of reconciling any number of rules against one another, determining the appropriate hierarchical relationship between rules, weighing precedence, considering past cases, and rendering a judgment that strikes an appropriate balance between potentially counter-vailing considerations. In some cases, it appears to be impossible to apply a law in isolation from its broader legal and jurisprudential contexts. For instance, the law established as Article 23, Par. (b) of the 1907 Hague Regulations states:

> It is especially prohibited...to kill or wound treacherously individuals belonging to the hostile nation or army.[20]

It might strike those unfamiliar with the conventional legal interpretation of this law as surprising that it has generally been interpreted as a prohibition against assassination whenever a state of war exists between parties. Louis Beres notes that there is a direct reference to Article 23(b) in Article 31 of the U.S. Army Field Manual 27-10, The Law of Land Warfare, dated 1956, which reads

> This article is construed as prohibiting assassination, proscription or outlawry of an enemy, or putting a price upon an enemy's head, as well as offering a reward for an enemy "dead or alive."[21]

Beres also notes that this understanding of Article 23 was entered into customary international law as of 1939, suggesting something of a resolution, but also introducing a further complication for legal rule-following; namely, there is also a temporal component in that the interpretation of the specific content of the law is prone to change over time. Still, the point remains that Article 23 stands as a paradigm case of the interdependence of laws in governing conduct in war, and probably in governing conduct in most other legal contexts as well.[22]

[19] Kevin Schieman, "The Soldier's Share: Considering Narrow Responsibility for Lethal Autonomous Weapons," *Journal of Military Ethics* 21/3–4 (2022), 228–245, here p. 239.

[20] "Convention with Respect to the Laws and Customs of War on Land," *The American Journal of International Law* 1/2 (April 1907 Supplement), 129-159, here p. 142.

[21] Louis Rene Beres, "After Osama Bin Laden: Assassination, Terrorism, War, and International Law," *Case Western Reserve Journal of International Law* 44/1 (2011), 93-147, here p. 109.

[22] It was brought to my attention during a presentation

A further difficulty for law-following is that legal contents may come into conflict with one another in ways that are conceptually difficult to resolve. Granted, law is generally hierarchical, but it is not always evident which guidance ought to take precedence in a particular case. For example, Convention I of the 1949 Geneva Convention, which addresses "the Amelioration of the Condition of the Wounded and Sick in Armed Forces in the Field." Convention I, Article 38 establishes the red cross and red crescent as markings identifying medical personnel and vehicles as prohibited targets and even prohibiting their improper use. However, it has not been uncommon during counter-insurgency operations to allow for the engagement of unmarked civilian vehicles, even where those vehicles appear to be providing life-saving medical care. This means that rules of engagement have often allowed soldiers to engage vehicles providing medical care on the legally justifiable grounds that they were not prohibited targets. Nevertheless, those vehicles would seem to satisfy the moral justification for the establishment of Article 38 in the first place (that is, the amelioration of the condition of the wounded and sick in armed forces in the field). It seems clear that reconciling a law that establishes protected status for medical personnel in war with a local rule allowing soldiers to target unmarked vehicles that may be performing a medical function depends on more than just the laws. It also depends on the context surrounding the application of those laws, a fact which applies to the law in general, but seems especially germane to resolving military law. The practical reality is that unmarked vehicles have often been used tactically to recover valuable weapons and intelligence from engagement sites, meaning that the rules of engagement—as they always do—reflect an effort to strike a balance between mission considerations and humanitarian concerns. Whether the rules of engagement strike the appropriate balance between those considerations is always a question worth entertaining—to quote John Austin, "the existence of law is one thing; its merit or demerit is another"[23]—but, the difficulty for law-following is that those considerations may be weighted differently at different levels of legal authority.

All of this serves simply to emphasize the more general point that following the laws of war depends on much more than any law in isolation. To the contrary, following the law requires an agent to reason prospectively about how the rules obligate and constrain legal subjects, and to reconcile any given application of those rules with relevantly similar past cases, a kind of judgment that John Rawls describes as reflective equilibrium. Put another way, the law follower needs to strike an adequate balance between understanding of the law and understanding of specific applications of the law. The difficulty for LAWS is that this type of consistency depends on an agent's ability to recognize and apply different types of rules in a loose hierarchy, to navigate the semantic content of those rules, and to reconcile those rules against countless other rules and cases. These challenges may not be irresolvable at least to the extent that nature has resolved them in human beings, but from the perspective of artificial intelligence the challenges are nevertheless substantial.

The Compliance Argument

I have argued here that following the law is actually far more demanding than one might assume and as a result, the technological bar to lethal autonomy is likewise higher than many people might have supposed. However, there is a compelling objection to this way of thinking about LAWS that deserves some attention. Arkin's claim is that engineers morally ought to build LAWS that follow the law, but maybe I have interpreted his claim too literally. Maybe what one ought to care about for LAWS is merely that they act in compliance with the law. Put another way, what military commanders want is machines that act in accord with the law, regardless whether they are engaged in anything like explicit rule following or not. To insist on anything more than compliance, one might argue, is to miss the point of Arkin's argument altogether. This is an incredibly important objection given the recent success of data-driven methods in artificial intelligence, and I must admit to being at

on this topic at the International Society of Military Ethics Conference in 2022 that Article 37 of the Geneva Convention (1949) prohibiting acts of perfidy offers an equally compelling example. The interpretation of prohibitions on perfidy depends at least as heavily on other law as the prohibition on assassination. I suspect that there are many and possibly better or more impactful cases than the one that I have identified here.

[23] John Austin, *The Province of Jurisprudence Determined and The Uses of the Study of Jurisprudence*, London, UK: Weidenfeld and Nicolson 1954, p. 184.

least somewhat sympathetic to the position. If it were possible to build machines that comply with the laws of war without explicit reference to those laws, then all parties to a conflict might still be considerably better off than they are with human soldiers.

Without fully resolving the objection, I want to voice a philosophical concern about any approach that seeks to achieve legal compliance without explicit reference to the laws themselves. On his theory of rule-following, Boghossian claims that for an agent to be engaged in rule-following, the agent must accept the rule, must act in ways that conform with the rule, and in virtue of accepting the rule, the agent's action must be explicable and rational (*ER* 472). Among other things, Boghossian's theory precludes the possibility of accidental rule-following in a way that rises to the level of moral concern where compliance with the law of war is concerned. That is to say that as I sit at my computer typing this paragraph, I am acting in accordance with infinitely many possible rules. I am acting in accordance with a possible rule that requires me to wear shoes while using the computer and another that requires me to listen to music while sitting in my office. I am likewise following a rule preventing me from unjustifiably killing another human being. In practice though, despite being compliant with each of these rules, imagined or otherwise, I am not in fact following them. I am not following them for my acceptance or rejection of those rules neither explains nor rationalizes my behavior; the rules are not exerting influence in either permitting or restricting my behavior, and as such, those rules are not guiding my conduct. And it is worth noticing that criminal law seems deeply concerned with considerations of explicability and rationalizability; criminal law is not simply a matter of determining whether a defendant's actions conform with the law—the law is generally quite concerned with intent.

Now, suppose a machine were trained on data that captured some set of legally acceptable actions in war, but possessed no explicit means of representing the laws themselves. I take it that such a machine would learn to comply with the law in the same way that a model like ChatGPT learns language, coming to recognize patterns over some massive corpus of data. If that were the case, then the machine would not learn that it is always illegal to intentionally target civilians; at best, it would learn something closely resembling that law. I take there to be a vast moral difference between a rule that prohibits targeting civilians categorically and one that prohibits targeting civilians *ceteris paribus*. The reality is that the opacity of data-driven machine learning—the inability to understand how a trained model is producing the outputs that it does—means that one can never be confident that an architecture that learns rules implicitly has, in fact, learned the proper rules. This should lead one to wonder if an implicit rule-following machine is adequate to the agential demands of choosing who lives and who dies in war.

Relatedly, the ability to follow rules seems to depend in important ways on an agent's ability to reason prospectively, applying a law to new tasks or contexts. In the previous section, I argued that one difficulty in applying the law is extending one's understanding of the law to new cases, but it is difficult to see how one could satisfactorily do so without some explicit representation of the law itself. The reality of data-driven methods in artificial intelligence is that they are notoriously brittle; they are capable of learning an incredible number of patterns in any given data set, but struggle to extend those patterns to new tasks or to accommodate even small context shifts in their data. Put another way, these networks are effective enough in interpolating between known cases, but so far have struggled to extrapolate beyond their trained tasks and data; the more similar that a new case is to trained cases, the more likely the system is to classify that case correctly. If that is the case, then it is entirely possible that data-driven methods would be technically wanting for all but the narrowest of tasks. I suppose then that my skepticism about the adequacy of data-driven, implicit rule-following machines is as much technological as it is philosophical, but questions regarding the appropriate use of technology to extend human agency should never be subordinated to questions of what is, in fact, possible.

Not Even Stupid

I have argued here that following the law is actually quite difficult in practice and that the demands of building a machine that is able to follow the law may be far more ambitious than one would give it credit for being. Inadvertently, I may have also offered a partial explanation for the fact that soldiers' compliance with the laws of war has often suffered in comparison with the understandably high expectations imposed by governments regarding the protection of the general public. Even for human soldiers, who benefit from an incredible capacity for abstract learning, compositional and hierarchical thought, and robust behavioral flexibility, the demands of rule-following are substantial. The problems posed by following the law—the heterogeneity of laws, semantic variation, and the global reasoning problem—set an extraordinarily high bar to competence for lethal autonomous weapons and for machine ethics more generally. That is not to suggest that technology will not surpass these considerable milestones; it might even happen far earlier than one could reasonably expect from the present vantage point. After all, recent success in massively scaling transformers, a type of machine learning network especially adept at language generation tasks, should serve as a reminder that advances in artificial intelligence have often been unpredictable, moving in dizzying fits and starts. As these networks have grown in size by orders of magnitude, they have given some indication that novel cognitive abilities may emerge at scale. This will, of course, exacerbate existing difficulties caused by operators' inability to understand machine behavior, especially where lethal consequences are concerned. Whatever the case, it is important that belligerent states approach these technological advances with a clear recognition that the use of sophisticated machine weapons will, as Heather Roff once told me,[24] both affect and reflect human moral agency.

[24] Private phone conversation in October of 2021.