Philosophy of Science (2025), **00**, 1–18
doi:10.1017/psa.2024.65

ARTICLE

Construct Validity in Automated Counterterrorism Analysis

Adrian K. Yee

Lingnan University, Department of Philosophy, Hong Kong Catastrophic Risk Centre, Hong Kong
Email: adrianyee@ln.edu.hk

(Received 18 July 2024; revised 18 September 2024; accepted 05 November 2024)

Abstract

Governments and social scientists are increasingly developing machine learning methods to automate the process of identifying terrorists in real time and predict future attacks. However, current operationalizations of "terrorist'" in artificial intelligence are difficult to justify given three issues that remain neglected: insufficient construct legitimacy, insufficient criterion validity, and insufficient construct validity. I conclude that machine learning methods should be at most used for the identification of singular individuals deemed terrorists and not for identifying possible terrorists from some more general class, nor to predict terrorist attacks more broadly, given intolerably high risks that result from such approaches.

Introduction

Social scientists and government intelligence services are increasingly developing "machine learning models of counterterrorism" (MMCs) using various forms of artificial intelligence (AI). For instance, the Israeli Defense Forces is actively using machine learning using the *Lavender* software system to identify alleged Hamas operatives so as to gather intelligence and coordinate air strikes on them (Davies and McKernan 2024). Others such as Python (2020) and Krieg et al. (2022) construct predictive models of future terrorist attacks using historical data, physical geographic data on locations of attacks, and news data leading up to the time of the attacks. However, in constructing such models, a variety of constructs of terrorism are defined and used, each of which presents risks of privacy violation, mistaken identification, unreliable predictive accuracy, and questionable explanatory power. And yet, there is almost no philosophical discussion on MMCs, with the exception of Verhlest et al. (2020) who argue that contemporary MMCs often face three methodological issues: an insufficient balance of positive and negative labels in training data leading to overfitting, an impractically vast number and kind of datasets required to predict terrorism above random chance, and the frequent positing of spurious correlations that intrinsically arise from analyzing high-dimensional data.

© The Author(s), 2025. Published by Cambridge University Press on behalf of the Philosophy of Science Association. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

In this article, I connect the philosophy of science literature on construct validity to counterterrorism studies and argue that most MMC methods are largely inadequate given three additional kinds of problems: insufficient construct legitimacy, insufficient criterion validity, and insufficient construct validity. I conclude that current MMCs generate models with far weaker predictive powers, and pose greater risks of harm, than has been previously acknowledged.

1. Automating counterterrorism analysis

1.1 Defining terrorism

Terrorism is a theoretical construct posited in many social scientific models as a means of describing a family of phenomena related to politically motivated violence. However, what counts as an adequate application of the term is unclear and often generates methodological confusion given that there remains no agreed upon definition of terrorism that is consistent between countries' governments, between agencies within a single country, nor even amongst social scientists. Currently, the closest we have is many scholars and governments deferring to the Global Terrorism Database's (GTD) definition: "the threatened or actual use of illegal force and violence by a non-state actor to attain a political, economic, religious, or social goal through fear, coercion, or intimidation" (START 2022). However, "in many cases agencies in the same country have adopted unique definitions" (LaFree and Dugan 2013, 4) illustrating how convergence of definition is difficult to obtain even when the stakes are high given the nature of terrorism. A recent systematic review of more than 100 definitions suggested that "the production of violence" is, at most, the sole unifying thread amongst extant definitions (Schmid 2023). This suggests that "terrorism" ought either to be considered a family resemblance term, or a term that should be broken up into other more fine-grained concepts with distinct applications depending on the specific purposes and values governments and social scientists may have in a given instance.

To illustrate, people have been arrested in the United States as terrorists for intending to provide medical aid to members of a paradigmatic terrorist organization (i.e., al-Qaeda), even if they are not members of such organizations and did not provide such aid (Barria 2005). Many state actors, for example Israel, have historically engaged in targeted assassinations against foreign adversaries to induce fear in their enemies (Bergman 2018), illustrating how some have applied the label to state actors. Indeed, during the French Revolution period, the Robespierre government was the first instance in which the term terrorism was applied, despite it being a state actor against its very own people (Hoffman 2017, 4). The Saudi Arabian government even once claimed in a 2014 antiterrorism law that anyone "[c]alling for atheist thought in any form, or calling into question the fundamentals of the Islamic religion on which this country is based is a terrorist" (Schmid 2023, 18), showing that terrorism has been used to describe epistemic threats to society. Many former US presidents have spoken of "terrorism" in divergent ways: Lyndon B. Johnson used the term to describe the Vietcong to appeal to South Vietnamese people, Ronald Reagan claimed that terrorists were the "anti-thesis to democracy" even if political violence and rioting is sometimes committed in the pursuit of democratic end goals (e.g., Hong Kong protestors circa 2019–20), and Bill Clinton claimed that terrorists were "enemies of

peace," even if terrorists are sometimes fighting with the intent to bring peace to their homelands (e.g., the Liberation Tigers of Tamil Eelam in Sri Lanka) (Gascón 2023). Such anarchic usage of the term "terrorism" may be somewhat unsurprising given that Stampnitzky (2013) provides evidence that academic social scientists in the United States did not seriously discuss the concept of terrorism until the 1970s, as measured by publications, conferences, and workshops using the term or adjacent concepts. Hence, most social scientific work remains a mere half-century old, despite related phenomena being ancient, with foundational methodological issues remaining to the present day.

Indeed, even when ad hoc definitions are used in studies, scholars admit that it is remarkably difficult to find reliable patterns concerning terrorism. Python 2020 (7) suggests four core features of terrorism: political aim, fear, publicity, and attacks on civilian targets, thereby differentiating the phenomenon from war, riots, robberies, or spontaneous acts of violence. However, history presents innumerable counter-examples to each of these features, leading some historians of terrorism to conclude that necessary and sufficient conditions will not be found (Schwenkenbecher 2012; Hoffman 2017). Furthermore, using data from 1968 to 2002, Piazza (2007) argues that there is essentially no robust correlation between socioeconomic factors such as poverty, malnutrition, inequality, unemployment, inflation, and poor economic growth, and participation in acts of terrorism. Additionally, against a common myth, at least two scholars conducting systematic reviews concluded that having higher education is positively correlated with joining terrorist organizations (Krueger 2007, xvi; Hoffman 2017, 173). While there is some evidence that variables such as ethnic and religious diversity, and the extent of state repression, are predictors of membership in terrorist groups (Ghatak and Gold 2017), others disagree and emphasize that such variables do not explain how many terrorists are raised in politically stable Western developed nations from sometimes affluent backgrounds (Dawson 2014). The lack of consensus on what factors affect membership in terrorist organizations had led one expert studying suicide terrorist attacks to write: "The only operational requirement is the ability to recruit persons with a willingness to kill and a willingness to die" (Hoffman 2017, 173). And yet, even this is not enough to enable experts to converge on a definition, nor on what counts as facts about covariates and their relationship to terrorist activities. The empirical social science on terrorism therefore remains nascent and unclear with respect to definitive conclusions.

Despite divergence of definitions on terrorism, philosophers of social science continue to emphasize that divergence of definitions can sometimes be a methodological virtue because it can ensure value pluralism and that relevant stakeholders' preferences are reflected in measurement constructs, such as what counts as misinformation (Yee 2023), adjudicating measures of democracy (Crasnow 2021), and metrics of well-being (Alexandrova and Haybron 2016). While this is granted, I will argue that the construction of MMCs presents a methodological challenge for value-pluralism about measurement in two senses. Firstly, adequate MMC development requires *precise* definitions for automated procedures to be conducted without error, even if disagreement is sometimes legitimate and fruitful about how to render the term precise. Secondly, terrorism is sometimes a global phenomenon requiring international coordination across nation state boundaries,

and thus requires at the very least convergence of agreement on ad hoc definitions for conducting counterterrorism operations. This is particularly manifested in the conduct of recent US military counterterrorism operations, in which coordination with foreign governments is required to produce methodologically adequate MMCs that aggregate large amounts of disparate data and diverging value judgments. As I will illustrate in the context of MMCs, convergence rarely happens in practice, given lack of agreement on what terrorism is, epistemic issues in obtaining reliable knowledge in intelligence analysis, and fundamental disagreements about political values that feature in judgments of terrorism.

1.2 Epistemic workflow in contemporary US counterterrorism

Despite aforementioned definitional concerns, government intelligence agencies and social scientists persist in developing computer-assisted methods for analyzing terrorism, in particular MMCs. There are broadly two kinds of MMCs currently developed: classifiers identifying both known and potential terrorists in real time using computer vision and natural language processing (Interpol 2024), and statistical models attempting to predict when an attack will happen in the future (Python 2020). I begin by describing both methods as they have appeared in recent US military applications. To understand the *epistemic workflow* (i.e., the sequence of knowledge-generating procedures) of MMCs, it is important to first describe the original manual procedures that are either presently automated or are intended to be automated in the future. My present focus will be on procedures conducted using contemporary unmanned aerial vehicle (drone) strikes used in the targeted killings of purported terrorists.

Project Maven began in 2017 as a computer vision research program within the US military aiming to enhance drones with the capability of automatically identifying combatants during the Defeat-ISIS campaign (Work 2017). This is part of a broader goal the US Department of Defense (DoD) has to automate weapons and military intelligence methods more generally: Scientific and engineering progress has reached the stage that automated target identification is intended to be conducted with sufficient autonomy that human operators may soon be wholly absent from weapons systems (Kechagias-Stamatis et al. 2017). Given the context of their usage, US military drone strike methodology is highly complex and classified, with little information publicly available. However, a whistleblower leaked classified information to the publication *The Intercept*, shedding light on the following legal and epistemic procedures used to select targets in countries such as Somalia and Yemen from 2011 to 2013, a geographic area in which US designated terrorist groups Al-Qaeda and Al-Shabaab continue to operate to this day (The Intercept 2015). I use this information as a case study to highlight and critique various methodological features of US military counterterrorism operations and MMCs.

Firstly, a typical drone strike would begin with Joint Special Operations Command Task Force 48-4 (specializing in Somalia and Yemen) working with the National Security Agency (NSA) and other DoD departments to construct what is called a "baseball card": a collection of known information about the potential target by drawing upon signals intelligence, statistical analysis, and other sources of information, especially intercepted phone calls, text messages, e-mails, and computer drives collected by field agents. For instance, the metadata from more than 55 million

phones in Pakistan alone have been tracked by the NSA and input into "a machine learning algorithm, which identifies likely couriers taking messages between likely terrorists. The algorithm is fed the mobile metadata on a small number of 'known terrorists.' It then sifts through the metadata of the rest to try to identify patterns that match those of the training set" (Danchev 2016, 705).

Secondly, while nearly all drone strikes are operated by personnel far from the regions in which they occur (e.g., US bases in Djibouti), most drone strikes practically require information gathering obtained from at least two kinds of personnel in the *local* geographical area of the target: on-the-ground US agents or their allies in local communities who conduct reconnaissance and select targets. However, when constructing MMCs, researchers have noted that while fieldwork is typically required to have appropriate data to train algorithms, many international counterterrorism researchers fail to obtain the relevant quantity or quality of fieldwork data (Atran et al. 2017). Indeed, whistleblowers have emphasized the ethical significance of the epistemological uncertainties introduced by these fraught procedures (The Intercept 2015):

> "There's countless instances where I've come across intelligence that was faulty." This, he said, is a primary factor in the killing of civilians. "It's stunning the number of instances when selectors are misattributed to certain people. And it isn't until several months or years later that you all of a sudden realize that it was his mother's phone the whole time."

Thirdly, a committee of US officials would approve strikes, being ultimately approved by the US president, and that the "country's government was also supposed to sign off." In the case of at least Somalia, "Somalia's minister of national security told *The Intercept* that the United States alerted Somalia's president and foreign minister of strikes .[H]e was unaware of an instance where Somali officials had objected to a strike, but added that if they did, he assumed the U.S. would respect Somalia's sovereignty" (The Intercept 2015). This process is intended to lend some degree of local political legitimacy to operations, though it remains unclear to what extent drone strikes are supported by locals. For instance, though a 2009 Gallup poll claimed only 9 percent of Pakistanis surveyed support drone strikes in their region (Bergen and Tiedeman 2011, 14), recent research suggests some support such strikes as defeating unwanted terrorist organizations that could not be accomplished through any other practical means by local governments (Ansari 2022). Hence, whether someone is considered a terrorist depends intrinsically upon background political value judgments by relevant stakeholders. In the words of former CIA director Leon Panetta, drones have often been perceived by the US government as "the only game in town" as a means of not only thwarting attacks against US and NATO forces operative during the war in Afghanistan but also for allowing the United States to defeat al-Qaeda and its allies to this day (Bergen and Tiedeman 2011, 17). Irrespective of whether one thinks such actions are morally justified, the US military continues to use drones as the default means for engaging adversaries, conducting reconnaissance, and to cooperate with other governments to conduct foreign policy (Fuller 2017). This is so despite their controversial political legitimacy and risks of harming innocent people, with the epistemic workflow for terrorist identification to be automated in the near future using MMCs and related software systems (Manson 2024).

This historical discussion elicits several salient ethical and epistemological features. Firstly, there has often been considerable difficulty retaining consistent identification of targets when conducting drone strikes. For instance, the minimum number of reconnaissance aircraft required to ensure that targets were visually identified correctly was often found lacking in the Somalia and Yemen regions of the Obama administration's drone activities circa 2011–15 (Currier and Maas 2015). Also, the entire epistemic network that military personnel have created for constructing "baseball cards" is acknowledged to be unreliable and is exacerbated by a lower bound estimate of more than 300 terabytes of terrorism related evidence that the US military has maintained for counterterrorism operations alone, despite lack of uniform procedures for analyzing this data (Fenzel et al. 2020). Nonetheless, the US military has used the SKYNET algorithm to conduct CIA drone strikes in Pakistan when gathering and interpreting phone data, in an effort to calculate the probability someone is a suspected terrorist, according to their records of known terrorists (Emery 2020, 3). This is so despite the fact that the US government has explicitly acknowledged between January 2011 and June 2012 alone that there was often lack of resources to positively identify striked targets with the standard of "near certainty" required for operations (The Intercept 2015).

As a means of mitigating these issues, machine learning is increasingly used in military intelligence analysis to supplant human error, and yet with controversial and unexpected outcomes. As I will argue, fully automated intelligence analysis systems are not currently ready for ethical deployment in counterterrorism applications. Indeed, we ought to have low confidence in the prospects of machine learning adequately automating epistemic workflow given that a recent survey of military intelligence analysis conducted manually by humans claims that "most substantive intelligence is not fact but expert judgment made under uncertainty" (Mandel and Irwin 2021, 559). Furthermore, the Combating Terrorism Center, the primary US military organization studying terrorism, has recently explicitly acknowledged these epistemological challenges in intelligence gathering when interfacing with AI systems (Rassler 2021).

Secondly, rendering epistemic workflow autonomous requires adequate operationalizations of "combatant," "terrorist," and other related concepts into AI systems. For instance, the most important definition of "international terrorism" in the US government appears[1] in "18 US Chapter Code 113B – Terrorism" section 2331 and concerns activities that satisfy each of conditions (A) - (C):

> (A) involve violent acts or acts dangerous to human life that are a violation of the criminal laws of the United States or of any State, or that would be a criminal violation if committed within the jurisdiction of the United States or of any State (B) appear to be intended to intimidate or coerce a civilian population to affect the conduct of a government by mass destruction, assassination, or kidnapping; and (C) occur primarily outside the territorial jurisdiction of the United States, or transcend national boundaries in terms of the means by which they are accomplished.

[1] https://uscode.house.gov/view.xhtml?path=/prelim@title18/part1/chapter113B&edition=prelim. Accessed April 1, 2024.

Given the logical structure of the definition, one can therefore deny the definition's adequacy by showing that one of its three conjuncts is unjustified in some manner. Indeed, the first conjunct (A) is problematic because it simply exports the laws of the United States to other countries when deciding who counts as a terrorist, where those in other countries may have good reasons to reject the specific laws of the United States as de facto structured to prevent paradigmatic cases of terrorism so defined.

Regarding part (B) of the law, there is also the fact that the boundaries of what counts as complicity in terrorism are sometimes ill-defined. For instance, it is unclear that someone who commits terrorism is the same as someone who conspires to do so but in such a manner that it is either not a sincere attempt or not likely to cause actual harm. Consider the conviction of al-Qaeda affiliate Kamel Bourgess for possessing castor beans allegedly used to make the biological toxin ricin, despite him being demonstrably incompetent and ricin not being a contagious chemical capable of mass casualties (Salama and Hansell 2005, 622). The point is that it is unclear how one could adequately automate the identification of prospective terrorists given that machine learning procedures are only as effective as is the adequacy of the measurement constructs and input data with which it is programmed. If we human beings struggle to make clear judgments pertaining to terrorism that experts can agree on, there is little hope for automating such procedures in machine code.

Regarding part (C) of the law, it remains unclear to what extent the US government can legally kill their own citizens in drone strikes, even if such citizens are outside US territory and call for violence, despite not committing violence themselves. Modern drone strike operations remain both conceptually and legally vague between being what are often either acts of undeclared and unofficial warfare or simply as a form of clandestine vigilante justice and international policing, thereby allowing the US government to frequently exercise arbitrary power in the pursuit of controversial targeted killings (Enemark 2021, 74–92). For instance, Ramsden (2011) discusses whether the US drone strike killing of al-Qaeda affiliate Anwar Al-Awlaki in 2011 is justified under International Human Rights Law, given he was a US citizen but that the strike was conducted outside of a declared war zone (i.e., Yemen). There are further cases in which citizens of non-US countries have had their citizenship removed and then executed, such as Bilal el-Berjawi in 2012 who was stripped of his British citizenship by the British government before being killed by a drone strike (The Intercept 2015). This has led legal scholars like Brooks (2014, 83) to declare that "drone strikes constitute a serious, sustained, and visible assault on the generally accepted *meaning* of certain core legal concepts, including 'self defense,' 'necessity', 'proportionality', 'combatant', 'civilian'."

The conclusion to draw here is that operationalizing what counts as a "terrorist" in MMCs will simply reinforce the biases of those involved in labeling the training data in possibly harmful ways that are unlikely to either satisfy the political objectives of the US DoD or be conducted ethically and adequately. As argued by Shoker (2021, 1), during the Obama administration pre-2012, men and boys (so-called military-age males) killed by drones were routinely excluded from collateral damage counts so as to assign combatant status by default and justify their deaths. As another example, the NSA used the SKYNET algorithm during operations in Yemen and Pakistan: "The program collects metadata and stores it on NSA cloud servers, extracts relevant

information, and then applies machine learning to identify leads for a targeted campaign" (Grothoff and Porup 2016). According to leaked slides from US military sources, the known false positive rate of 0.18 percent is very high considering that the base rate is also high. This is critical to note given that many drone strikes are executed on mere "signatures" of a purported terrorist's presence, as estimated from computer vision outputs of suspects' physical appearance and cell phone signals analyzed with natural language processing (The Intercept 2015). In such cases, strikes occur even absent direct and firm evidence for a target being accurately identified, and are instances where AI systems have been and will continue to be used to conduct signature strikes, especially in light of Project Maven's computer vision research program. This is so despite it being implausible that such current procedures for defining "terrorist" could be automated with AI to the degree required for ethical usage.

Thirdly, even assuming that definitions of "terrorist" could be adequately operationalized in MMCs, recent military psychologists argue that even experienced intelligence analysts lack shared semantics for the dissemination and interpretation of reports, which can cause serious risks and actual harm to a variety of stakeholders. For example, Irwin and Mandel (2023) report that both military intelligence analysts and laypeople routinely incorrectly conflate "probabilities" with "confidence," that around 25 percent of experts and as much as more than 50 percent of nonexperts provided inconsistent numerical translations of graded credence statements such as "likely," and that analysts prefer information presented as graded credences equally as much as they prefer specific numerical probabilities, thereby causing considerable confusion in communicating intelligence. By some estimates, up to 82 personnel and technical support staff can be involved in the execution of a single drone strike, such as for the popular Predator class of drones used since the 1990s (Marra and McNeil 2013, 1168). Given that these personnel are liable to make human errors in gathering, analyzing, and communicating military intelligence, it is therefore unsurprising that, for instance, 90 percent of the more than 200 targets killed by US drones in Operation Haymaker (2012–13) in Afghanistan were not the intended targets (The Intercept 2015). The prospects are therefore poor for automating counterterrorism procedures using machine learning if human beings already struggle to agree upon definitions and methods of communication, especially given disagreements on value judgments.

2 Construct legitimacy, criterion validity, and construct validity

The core philosophical issues present in the previous historical discussion can be summarized as resulting from three kinds of neglected problems: construct legitimacy, criterion validity, and construct validity. Following recent philosophy of science, *construct legitimacy* concerns whether a construct *C* of terrorism is justified as evaluated by background social scientific virtues (Stone 2019). By way of contrast, *criterion validity* concerns whether an MMC produces outputs that cohere with other MMCs and metrics of terrorism in ways that are deemed satisfactory by background theory (Zhao 2023). Lastly, *construct validity* of an MMC concerns whether an MMC that measures *C* adequately tracks *C* (Alexandrova and Haybron 2016).

2.1 The problem of construct legitimacy

I begin with a discussion of construct legitimacy prior to discussions of criterion validity and construct validity. Operationalizations of terrorism in many MMCs are not construct legitimate for the reason that terrorism is arguably not a unified phenomenon: The term does not have enough referential clarity nor stability over time for linguistic participants to understand what it means in the majority of cases. As the previous historical discussion has illustrated, lack of consistent operationalization risks MMCs producing outputs that are alien to their users and designers, introducing risks of unintended outputs.

Additionally, the preceding discussion demonstrates how "terrorism" is an intrinsically value-laden construct that contains both descriptive and normative components. The purpose of positing terrorism as a construct is that it is intended to function as a conceptual intermediary in some unobservable latent space whose properties we infer from observable data (e.g., humans engaged in violent events) and whose existence is supposed to explain the data. That is, if one was interested to understand why some people commit paradigmatic terrorist acts, the thought is that it is because all such people are involved in "terrorism" as a sufficiently unified phenomena that is worthy of being referred to as a singular construct. Historically, the existence of a construct in many social sciences was posited as a result of "inference to the best explanation" reasoning, such as how psychometricians posited constructs of general intelligence (e.g., Spearman's *g*) as a way to explain how children make accurate judgments across sensory modalities (Zhao 2023). MMC researchers have implicitly followed an analogous methodology in constructing models attempting to identify paradigmatic terrorists and predict prototypical terrorist attacks while employing the construct of terrorism (National Institute of Justice 2023).

The significance of this for the problem of construct legitimacy is that terrorism is not an observable term in the same way that an electron's mass or someone's height is an observable term: One and the same violent action can be considered a terrorist action or not depending on the observer's background values. To see this, notice that what counts as someone's height is fully determined by intersubjectively verifiable properties of one's physical distance as measured from head to toe. By way of contrast, what counts as a terrorist act depends entirely on whether one judges the action as having either some degree of legitimacy, whether one finds the action constitutive of a significant enough amount and kind of violence (e.g., a cyberattack is not a terrorist act in the same way a car bombing is), or whether terrorist attacks can only apply to innocent people, among other factors featuring in value judgments related to terrorism. Therefore, considerations of empirical data alone never determine whether one is a terrorist. This entails that value judgments must either be directly inputted into MMCs as parameters in models or that value judgments are implicitly contained in feature extraction using training data. And yet, doing so is rarely feasible in a manner that accurately preserves the input human values over time given that there is already lack of clarity in the human context prior to operationalization.

To clarify, this is not to say that there cannot be any construct legitimate applications of the term terrorism in any social scientific model; there are clearly

justified usages of such constructs, especially in specific and narrower contexts of usage (e.g., identifying known al-Qaeda operatives who have been proven to have committed violent acts). The point is rather that in the act of automating the identification of terrorists in machine learning, the problem is far more salient in MMCs than in other kinds of counterterrorism contexts. This is so given that MMCs simply reinforce extant biases of the human analysts designing and training them, in ways that are sometimes known but also sometimes unknown by their creators. Because value judgments are made by a variety of stakeholders, some of which have disproportionate amounts of power in their societies compared to others (e.g., politicians with diverging interests from their citizens), disagreement can lead to highly contested definitions of terrorism that risk generating MMCs that produce counterintuitive or unjustified outputs, leading to even innocent people being killed.

To illustrate more explicitly how value judgments can play a role in MMCs and violate construct legitimacy, Python (2020) argues in a book-length monograph on MMCs that the following seven claims about terrorism are theoretically inadequate or empirically false for the period 2002–17: terrorism is well defined, terrorism aims only at killing civilians, Western nations are particularly vulnerable to terrorism, terrorism is increasing over time, terrorism occurs randomly, locations of terrorism are static, and terrorism cannot be predicted. Python argues that given how fraught the social science of terrorism is, terrorism has not been defined similarly enough across databases that it makes cross-study comparisons and systematic reviews nearly impossible to conduct: "Finding the most frequent words does not suffice to account for the diversity of views on terrorism. Several equally valuable views on terrorism may coexist and some degree of subjectivity in the interpretation of the concept of terrorism cannot be avoided" (ibid., 6–7). For instance, the Democratic Republic of Congo witnessed nearly twice as many instances of state actors killing civilians than nonstate actors, suggesting that whether a "terrorist" is defined as a nonstate actor produces different results from MMCs (ibid., 16). Hence, different definitions arising from differences of value judgments imply radically different conclusions one can draw from models (e.g., whether states can be terrorists or only nonstate actors).

The philosophical upshot is that MMC theorists may be better off acknowledging that there are multiple different kinds of phenomena (e.g., different forms of violence) being measured that are being confused for one unified phenomenon (i.e., terrorism). If this is right, terrorism could be better understood as a *ballung* concept: concepts that admit of equally valid but distinct measures given differences in the aims, values, and scope of stakeholders and policy makers (Cartwright and Runhardt 2014). But the problem of construct legitimacy still remains even given multidimensional metrics: Relevant stakeholders (e.g., citizens, politicians, and social scientists) may still disagree on whether a specific operationalization of "terrorist" is adequate if there is significant divergence of value judgments concerning what counts as a terrorist act. Thus an MMC's construct legitimacy is contingent upon the extent at which there is enough agreement on the definition being operationalized in machine code amongst relevant stakeholders. This is especially so when one considers recent cases in international law and human rights, where targeted killings using drone strikes are often conducted by, for example, the US military in countries whose citizens (a core stakeholder in definitions of "terrorist") are either often mistakenly

killed or whose citizens do not fully consent to operations in those countries (e.g., Pakistan), thus rendering the construct illegitimate (Bergen and Tiedeman 2011).

2.2 The problem of criterion validity

The problem of construct legitimacy is to be contrasted with the problem of *criterion validity* (Zhao 2023). In the case of criterion validity, we are interested in whether an MMC is producing outputs that correlate coherently with the outputs of other measures related to terrorism that are already deemed satisfactory by various background methodological criteria. For example, in psychometrics, if someone is designing an intelligence test by creating two subtests for mathematical and spatial reasoning, whenever a subject scores highly on both, then both tests are mutually criterion valid for producing outputs that are coherent with one another, given that the capacity for spatial and mathematical reasoning are plausibly correlated with intelligence more general.

To illustrate the way in which many MMCs suffer from a problem of criterion validity, consider the multilevel analysis conducted by Jiang et al. (2023). Using data from the GTD, they construct two models, each of which draws upon three levels of features: microscopic (e.g., individual terrorist characteristics), mesoscopic (e.g., connections between various terrorist groups), and macroscopic (e.g., nationwide and international political issues). Using data ranging from the specific terrorist group, the attack's location, nationality of terrorist members, choice of weaponry, method of attack, and type of target, they gathered satellite data to construct a spatial grid of countries around the world determined by this data. The task was to predict the location of an attack by the Pakistan-based terrorist organization Tehrik-i-Taliban Pakistan at a particular time given data available previous to that time. Similar studies such as Khan et al. (2023) argue that mainstream algorithmic methods in MMCs are so effective at predicting terrorist attacks that they claim to have constructed a classifier with 95 percent accuracy for predicting the type of weapon used in an attack, as evaluated with respect to incidents in India, Pakistan, and Afghanistan.

Despite ostensible empirical success, a problem of criterion validity can arise in the following sense. In the context of MMCs, two models are mutually criterion valid if they cohere with one another and both produce measures concerning terrorist activity that are both respectively construct legitimate on their own terms. The problem here is that different MMC models can easily disagree with one another as to what are the relevant features of a person that are correlated with their propensity to commit terrorist attacks, even when terrorism is defined in the same way. For instance, just like Khan et al. (2023), Python (2020) also used similar kinds of satellite data to correlate physical locations of terrorism at one time with future attacks, to study attacks in Iraq, Afghanistan, and Pakistan from 2002–17, producing a model resulting in a false positive rate as high as 21 percent (ibid., 108). Here, one and the same method of measurement (i.e., satellite data) used by two distinct MMCs developed by different researchers is producing distinct and unreliable outputs, with drastically different rates of empirical success, despite using the same definition of terrorist.

To make matters worse, nearly 50 percent of the GTD's documented global terrorist attacks committed from 2002–19 are by *unknown* actors with unknown characteristics (Jiang et al. 2023, 14). But if nearly 50 percent of terrorist attacks in

this period are committed by actors whose features we know *nothing* about (e.g., their gender, age, ethnicity, income, education), this suggests that the existing GTD data most scholars use to train MMC models is not sufficiently representative of the true distribution of features that are relevant covariates connected to terrorist attacks, activities, and membership. Criterion validity is therefore lacking insofar as some social scientists apply the same construct of terrorism to ostensibly similar classes of violent events despite often having no clear idea of the relevant kinds of features terrorists possess such that researchers can ensure whether measures cohere with one another. Indeed, the previous historical discussion has shown no clear patterns between features determining whether an individual is likely to join a terrorist organization or commit an act of violent terrorism. This leads to not only epistemologically fragmented models but also models whose criterion validity is at best obscure and at worse wholly invalid. Hence, researchers must first agree on what are the relevant kinds of features that pertain to paradigmatic terrorist attacks if they are to ensure criterion validity and make accurate predictions.

2.3 The problem of construct validity

Criterion validity is to be contrasted with *construct validity* in the following sense: A measure *M* of a construct *C* is construct valid whenever *M* successfully and reliably tracks all the properties of *C* that *M* was designed to track (Alexandrova and Haybron 2016).

The problem of construct validity is particularly manifested in the context of MMCs in the temporal dimension, or what Michel (forthcoming) calls *validity drift* in psychometrics: Researchers sometimes end up measuring something very different at later points in time than what they originally intended to measure. Relatedly, Tal (2019) notes that when multiple measurement procedures designed to measure the same phenomenon disagree, it is unclear whether each procedure is independently inaccurate or whether each measures distinct concepts. One can apply this concept to temporal contexts in the sense that if an MMC fails to make accurate predictions about the future, it is unclear whether it is failing to predict terrorism or whether the relevant concept of terrorism has changed in a way that researchers fail to recognize that it has changed.

For example, imagine someone defining terrorism in the late-eighteenth-century French sense of only applying to violent political events committed by state actors (Hoffman 2017). If one trained a supervised learning MMC on data labeled with eighteenth-century historical cases of terrorism so defined, it would make radically false predictions about twentieth- and twenty-first-century terrorism because it fails to identify intuitive cases of present day terrorist attacks that are committed by nonstate actors. This renders moot any diachronic analysis of terrorism over time: Independent arguments must be provided that demonstrate that the construct is justifiably being used in the same way in different temporal contexts. In this hypothetical example, there is therefore a lack of construct validity in the sense of validity drift: The construct of terrorism that the MMC developed through supervised learning is not tracking the relevant features of these present-day terrorist attacks given the way it was trained. This case is not unrealistic insofar as there are many reasons to think that social scientists and governments continue to use different

definitions and employ distinct intuitive concepts of terrorism in their minds, in diverging ways over time (Schmid 2023). Given the aforementioned historical evidence, there are reasons to believe that validity drift can even occur in the short term, leading to poor predictive modeling, false positive targeted killings of purported terrorists, and other high-risk outcomes.

More generally, any social construct employed in machine learning ought to have sufficient stability of the semantics of that term such that measures adequately model the same phenomenon over time. Yee (2023) has argued that construct validity in machine learning models of *misinformation* suffer from an analogous problem to the case of MMCs: What counts as misinformation is not only irreducibly context sensitive but also changes drastically over time, even if misinformation is defined as "false or misleading information." This is because misinformation is arguably a value-laden concept where one and the same proposition uttered on a social media platform can be judged to be misinformation according to one set of values and not misinformation according to another. For instance, the claim that "COVID-19 lockdowns are bad" is misinformation for those who value lowering COVID-19 mortality rates and yet not misinformation for those who value economic growth and mental health, considering that lockdowns negatively impacted these latter features. Analogous considerations apply in the case of terrorism prediction: One and the same event can be considered terrorism by one group of relevant stakeholders judging a violent event and yet fail to be considered terrorism by another (e.g., whether Israel's usage of MMCs to bomb hospitals in Gaza is constitutive of terrorism or a legitimate military action against Hamas operatives allegedly hiding in these buildings). This further illustrates how whether a violent act counts as terrorism is not a property of the event but of the value judgments of observers. This context sensitivity therefore renders the construct validity of an MMC a function of the protean value judgments of relevant stakeholders over time, where such judgments can change at later points in time and inhibit construct validity.

Relatedly, Feest (2020) argues that construct validity comes in degrees: A test or measure can have construct validity relative to one coarse-grained feature of a construct and yet fail to be construct valid at finer-grained divisions. For example, in the case of MMCs, Krieg et al. (2022) criticize some MMCs for not dividing up geographic regions into finer grained categories (e.g., West Africa is taken as an entire unit of analysis in some MMCs, with different results occurring when individuated by country or subregions). This analysis applies more broadly to cases of the granularity of time variables in MMCs: "[T]o make meaningful predictions at a granular timescale (e.g., will a terrorist attack take place during a given week), models must have inputs of a similar temporal granularity in order to differentiate between points in time" (ibid., 2). Hence, the construct validity of MMCs is a function of the extent at which it is appropriate for the specific temporal context of deployment.

To close this discussion, construct validity is often important to retain insofar as, for example, NATO countries need to coordinate analytical taxonomies and procedures regarding terrorist threats across jurisdictions that individually use distinct taxa (Mandel and Irwin 2021). Usage of MMCs in shared analytical contexts cannot function reliably when multiple conflicting constructs are being used in the same mode of analysis: Machine learning is not the appropriate methodology for aggregating data from disparate taxonomies given the precision required of

computing systems. When international or cross-jurisdictional analyses are required, a lack of convergence on concepts of "terrorism" and their operationalizations in machine code has, as a matter of historical fact, often generated confusion and harmed innocent lives. It is nonetheless granted that the usage of MMCs can be permissible in restricted cases in which such systems have an epistemic workflow that is sufficiently isolated from other intelligence agencies' distinct taxonomies, thereby justifying their usage in their specific, independent use cases and respecting value pluralism about construct validity. The philosophical point is that value pluralism has its limits and that convergence of agreement on constructs is especially important when different stakeholders need to agree on how to define a construct, especially in machine learning contexts.

3 MMCs should only identify individuals, not members of classes

I have argued that there are three kinds of problems confronting contemporary MMCs. Given the current state of counterterrorism analysis in both US military applications and more general social science applications using machine learning, MMCs ought not to be used for predicting instances of a more general class of "terrorists," as determined using feature extraction procedures generated by MMCs, nor should they be used to try to forecast future terrorist attacks. Skepticism about the present and future prospects for most MMCs is warranted for reasons I have highlighted previously: Most military intelligence is known to be unreliable, sometimes internally inconsistent, and employs a wide variety of definitions such that false positive and false negative judgments are more likely to occur than not. The current state of the epistemology of contemporary counterterrorism is not sufficiently sophisticated nor robust to outliers such that construct legitimacy, criterion validity, and construct validity are satisfied to the degree necessary to avoid significant risks that arise from improper usage of MMCs. The stakes are high considering that MMCs are not a mere academic exercise in predictive social science but are actively used by many governments to actively seek and kill targets internationally on a near weekly basis. And yet, many methodological issues that already arise in the context of manual counterterrorism operations suggest that continuing to automate what used to be manual epistemic workflow procedures into MMCs simply exacerbates the underlying methodological issues of bias in feature selection, lack of relevant data, and high dimensionality.

Nonetheless, I do believe that it remains fruitful to continue MMC applications concerning the identification of *singular and known* terrorists only if a sufficient quantity and kind of data is available. While it is often not the case that such data is sufficiently available, rendering MMC usage even in these contexts less justified than some governments believe is the case, this latter kind of inference task is radically distinct from trying to identify terrorists as members of some generic class of people, or predict future terrorist attacks in general. Indeed, there is empirical evidence demonstrating such risks given that the Nexus 7 software tool, used in the 2000s by the US military to track known individual terrorists and terrorist groups operative in Iraq and Afghanistan, was in the best of cases only 70 percent accurate, with no independent oversight body to verify the construct legitimacy of this algorithm (González 2015, 16). However, the identification of singular individuals remains an

easier task to justify methodologically, given that comparatively less issues of construct legitimacy, criterion validity, or construct validity apply. This is because one does not need a general concept of what a terrorist is to identify specific individual people that are deemed terrorists. Rather, one simply needs either sufficient information on the person's appearance and behaviour or a working definition of "terrorist" in a specific legal and political context that can be justified enough to the relevant stakeholders to begin counterterrorism operations. For example, it is near universally agreed, at least outside Salafi Jihadist circles, that Osama Bin Laden was a paradigmatic example of a terrorist; one does not first need an antecedent general theory of terrorism to determine that he is a terrorist. This is because we know enough of the relevant facts about Bin Laden's activities to have judged that he is a "terrorist" relative to most countries' values.

However, one *does* need some theory of terrorism if one wishes to predict future terrorist attacks, especially in machine learning contexts, insofar as many MMC practitioners are interested in identifying future instances of a class of "terrorists" more generally. The reason is that one will require a supervised learning algorithm to have labeled data, where labeling data should be done in a methodologically rigorous fashion using a background theory of what counts as a terrorist, so as to avoid the objection that the algorithm is failing construct validity. Even in unsupervised learning contexts, where labeled data is not necessary, both the construct legitimacy and the construct validity of the MMC will be a function entirely of some background theory of terrorism that practitioners are implicitly committed to. Hence, the kinds of MMCs that are potentially justifiable in the present are those which use image recognition software (e.g., through convolutional neural networks) or those employing natural language processing to decipher intercepted phone calls so as to identify singular individuals who are known terrorists. While, for example, facial recognition tasks using MMCs continue to face serious methodological and ethical issues (Robbins 2021), such as systemic algorithmic bias, privacy violations, and falsely implicating a person in terrorist activity, such methods are at least not as susceptible to the same degree of methodological issues as the kind I have discussed. Nonetheless, more caution should be exercised concerning the usage of such software systems than social scientists and governments developing MMCs have expressed.

Conclusion

Philosophers of science have increasingly discussed issues of construct legitimacy, criterion validity, and construct validity in the context of social sciences such as psychology, psychiatry, well-being studies, and misinformation studies, amongst others. I have argued that counterterrorism studies presents another field in which a lack of clarity concerning measurement procedures can lead to harmful outcomes, especially as manifested in the context of the epistemic workflow for recent MMCs. Because counterterrorism operations are conducted by countries with differing values and purposes, and yet often need to coordinate on shared concepts of terrorism to successfully conduct international operations, failure to recognize the aforementioned problems of measurement has already led to false positive identification of terrorists, harm to innocent lives, and methodological inadequacy.

The broader philosophical conclusion is that while value pluralism about constructs is often warranted in local contexts of application, terrorism as an international phenomenon presents the challenge that sufficient agreement on definitions is nonetheless required to conduct multilateral joint operations globally, especially in machine learning applications. While it is beyond the scope of this article, it is possible that other socially constructed phenomena such as "protester" or "activist" may pose analogous challenges of measurement in machine learning contexts as well, thereby illustrating how the problems discussed here may generalize beyond counterterrorism. For now, I believe the empirical and historical record strongly suggests that given ongoing methodological challenges applications of MMCs should be restricted to, at most, the identification and analysis of singular individuals deemed terrorists.

Acknowledgments. I thank Andre Curtis-Trudel, Christopher Fuller, and Kenji Hayakawa for critical feedback on ideas in this paper. All errors, infelicities, and opinions are mine alone. I acknowledge funding from the Hong Kong Catastrophic Risk Centre and two Hong Kong government grants: the Research Matching Grant Scheme #185249 and the Faculty Research Grant #101914 both identically titled "Machine Learning Models of Misinformation and Deceptive Media."

References

Alexandrova, Anna, and Daniel M. Haybron. 2016. "Is construct validation valid?" *Philosophy of Science* 83:1098–1109. https://doi.org/10.1086/687941.

Ansari, Neha. 2022. "Precise and popular: Why people in northwest Pakistan support drones." *War on the Rocks*. https://warontherocks.com/2022/08/precise-and-popular-why-people-in-northwest-pakistan-support-drones/. Accessed February 23, 2024.

Atran, Scott, Robert Axelrod, Richard Davis, and Baruch Fischhoff. 2017. "Challenges in researching terrorism from the field." *Science* 355 (6323):352–54. https://doi.org/10.1126/science.aaj203.

Barria, Carlos. 2005. "Two US citizens charged with helping al-Qaida." *NBC News*. https://www.nbcnews.com/id/wbna8030379. Accessed June 29, 2024.

Bergen, Peter, and Katherine Tiedemann. 2011. "Washington's phantom war: The effects of the US drone program in Pakistan." *Foreign Affairs* 90 (4):12–18.

Bergman, Ronen. 2018. *Rise and Kill First: The Secret History of Israel's Targeted Assassinations*. New York, NY: Random House.

Brooks, Rosa. 2014. "Drones and the international rule of law." *Ethics & International Affairs* 28 (1):83–103. https://doi.org/10.1007/S0892679414000070.

Cartwright, Nancy, and Rosa Runhardt. 2014. "Measurement." In *Philosophy of Social Science: A New Introduction*, 265–287. Oxford: Oxford University Press.

Crasnow, Sharon. 2021. "Coherence objectivity and measurement: The example of democracy." *Synthese* 199:1207–29. https://doi.org/10.1007/s11229-020-02779-w.

Currier, Cora, and Peter Maas. 2015. "Firing blind: A secret Pentagon study highlights the chronic flaws in intelligence used for drone strikes in Yemen and Somalia." *The Intercept*. https://theintercept.com/drone-papers/firing-blind/. Accessed March 25, 2024.

Danchev, Alex. 2016. "Bug splat: The art of the drone." *International Affairs* 92 (3):703–13. https://doi.org/10.1111/1468-2346.12609.

Davies, Harry, and Bethan McKernan. 2024. "IDF colonel discusses 'data science magic powder' for locating terrorists." *The Guardian*. https://www.theguardian.com/world/2024/apr/11/idf-colonel-discusses-data-scienc E-magic-powder-for-locating-terrorists. Accessed June 29, 2024.

Dawson, Lorne. 2014. "Trying to make sense of home-grown radicalization." In *Religious Radicalization and Securitization in Canada and Beyond*, edited by Paul Bramadat and Lorne Dawson, 64–91. Toronto: University of Toronto Press.

Emery, John. 2020. "Probabilities towards death: Bugsplat, algorithmic assassinations, and ethical due care." *Critical Military Studies* 8 (2):179–197. https://doi.org/10.1080/23337486.2020.1809251.

Enemark, Christian. 2021. "Drone violence as wild justice: Executions on the terror frontier." In *Ethics of Drone Strikes: Restraining Remote-Controlled Killing*, edited by Christian Enemark, 74–92. Edinburgh: Edinburgh University Press.

Feest, Uljana. 2020. "Construct validity in psychological tests: The case of implicit social cognition." *European Journal for Philosophy of Science* 10 (4):1–24. https://doi.org/10.1007/s13194-019-0270-8.

Fenzel, Michael, Leslie Slootmaker, and Kim Cragin. 2020. "The strategic potential of collected exploitable material." *Joint Forces Quarterly* 99:31–39.

Fuller, Christopher J. 2017. *See It/Shoot It: The Secret History of the CIA's Lethal Drone Program*. New Haven, CT: Yale University Press.

Gascón, José Ángel. 2023. "The inferential meaning of controversial terms: The case of "terrorism."" *Topoi* 42:542–59. https://doi.org/10.1007/s11245-022-09879-x.

Ghatak, Sambuddha, and Aaron Gold. 2017. "Development, discrimination, and domestic terrorism." *Conflict Management and Peace Studies* 35 (6):618–39. https://doi.org/10.1177/0738894215608511.

González, Roberto J. 2015. "Seeing into hearts and minds: Part 2. 'Big data', algorithms, and computational counterinsurgency." *Anthropology Today* 31 (4):13–18. https://doi.org/10.1111/1467-8322.12188.

Grothoff, Christian, and J. M. Porup. 2016. "The NSA's SKYNET program may be killing thousands of innocent people." *Ars Technica*. https://arstechnica.com/information-technology/2016/02/the-nsas-skynet-program-may-be-killing-thousands-of-innocent-people/3/. Accessed March 29, 2024.

Hoffman, Bruce. 2017. *Inside Terrorism*. 3rd ed. New York: Columbia University Press.

Interpol. 2024. Identifying terrorist suspects. *Interpol.* https://www.interpol.int/Crimes/Terrorism/Identifying-terrorist-suspects. Accessed July 15, 2024.

Irwin, Daniel, and David Mandel. 2023. "Communicating uncertainty in national security intelligence: Expert and nonexpert interpretations of and preferences for verbal and numeric formats." *Risk Analysis* 43:943–57. https://doi.org/10.1111/risa.14009.

Jiang, Dong, Jiajie Wu, Fangyu Ding, Tobias Ide, Jürgen Scheffran, David Helman, Shize Zhang, Yushu Qian, Jingying Fu, Shuai Chen, Xiaolan Xie, Tian Ma, and Mengmeng Hao Quansheng Ge. 2023. "An integrated deep-learning and multi-level framework for understanding the behavior of terrorist groups." *Heliyon* 9 (8):1–17. https://doi.org/ 10.1016/j.heliyon.2023.e18895.

Kechagias-Stamatis, Odysseas, Nabil Aouf, and David Nam. 2017. "3D automatic target recognition for UAV platforms." *Sensor Signal Processing for Defense Conference*. 1–5. https://doi.org/10.1109/SSPD.2017.8233223.

Khan, Fahad Ali, Gang Li, Anam Nawaz Khan, Qazi Waqas Khan, Myriam Hadjouni, and Hela Elmannai. 2023. "AI-driven counter-terrorism: Enhancing global security through advanced predictive analytics." *IEEE Access* 11:135864–79. https://doi.org/10.1109/access.2023.3336811.

Krieg, Steven, Christian W. Smith, Rusha Chatterjee, and Nitesh V. Chawla. 2022. "Predicting terrorist attacks in the United States using localized news data." *PLoS ONE.* 17 (6):1–26. https://doi.org/10.1371/journal.pone.0270681.

Krueger, Alan. 2007. *What Makes a Terrorist*. Princeton: Princeton University Press.

LaFree, Gary, and Laura Dugan. 2013. "The global terrorism database: 1970–2010." In *Handbook of Computational Approaches to Counterterrorism*, edited by V. S. Subrahmanian, 3–22. New York: Springer.

Mandel, David R., and Daniel Irwin. 2021. "Uncertainty, intelligence, and national security decision making." *International Journal of Intelligence and Counter-Intelligence.* 34:558–82. https://doi.org/10.1080/08850607.2020.1809056.

Manson, Katrina. 2024. "US used AI to help find Middle East targets for airstrikes." *Bloomberg.* https://www.bloomberg.com/news/articles/2024-02-26/us-says-it-used-ai-to-help-find-targets-it-hit-in-iraq-syria-and-yemen. Accessed July 15, 2024.

Marra, William and Sonia Mcneil. 2013. "Understanding "The Loop": Regulating the Next Generation of War Machines." *Harv. J. L. & Pub. Pol'y* 36 (3):1140–85.

Michel, Matthias. Forthcoming. "Validity drifts in psychiatric research." *British Journal for the Philosophy of Science.* https://doi.org/10.1086/730534.

National Institute of Justice. 2023. "Research rooted in machine learning challenges conventional thinking about the pathways to violent extremism." https://nij.ojp.gov/topics/articles/research-rooted-machine-learning-challenges-conventional-thinking-about-pathways. Accessed July 15, 2024.

Piazza, James. 2007. "Rooted in Poverty?: Terrorism, Poor Economic Development, and Social Cleavages." *Terrorism and Political Violence* 18 (1):159–77.

Python, Andrew. 2020. *Debunking Seven Terrorism Myths Using Statistics*. New York: CRC Press.
Ramsden, Michael. 2011. "Targeted killings and international human rights law: The case of Anwar Al-Awlaki." *Journal of Conflict & Security Law* 16 (2):385–406. https://doi.org/10.1093/jcsl/krr015.
Rassler, Don. 2021. "Data, AI, and the future of US counterterrorism: Building an action plan." *CTC Sentinel* 14 (8):31–44.
Robbins, Scott. 2021. "Facial recognition for counter-terrorism: Neither a ban nor a free-for-all." In *Counter-Terrorism, Ethics and Technology*, edited by Adam Henschke, Alastair Reed, Scott Robbins, and Seumas Miller, 89–104. Cham: Springer.
Salama, Sammy, and Lydia Hansell. 2005. "Does intent equal capability? Al-Qaeda and weapons of mass destruction." *Nonproliferation Review* 12 (3):615–53. https://doi.org/10.1080/10736700600601236.
Schmid, Alex. 2023. "Defining terrorism." *International Centre for Counter-Terrorism*. https://www.icct.nl/sites/default/files/2023-03/Schmidt%20-%20Defining%20Terrorism_1.pdf. Accessed July 28, 2024.
Schwenkenbecher, Anne. 2012. *Terrorism: A Philosophical Inquiry*. London: Palgrave MacMillan.
Shoker, Sarah. 2021. *Military-Age Males in Counterinsurgency and Drone Warfare*. Cham: Palgrave Macmillan.
Stampnitzky, Lisa. 2013. *Disciplining Terror*. Cambridge: Cambridge University Press.
START (National Consortium for the Study of Terrorism and Responses to Terrorism). 2022. Global Terrorism Database 1970–2020. https://www.start.umd.edu/gtd. Accessed June 19, 2024.
Stone, Caroline. 2019. "A defense and definition of construct validity in psychology." *Philosophy of Science* 86:1250–61. https://doi.org/10.1086/705567.
Tal, Eran. 2019. "Individuating quantities." *Philosophical Studies* 176 (4):853–78. https://doi.org/10.1007/s11098-018-1216-2.
The Intercept. 2015. The Drone Papers. *The Intercept*. https://theintercept.com/drone-papers/. Accessed July 15, 2024.
US Department of Defense. 2022. "(U) evaluation of contract monitoring and management for Project Maven." *US Department of Defense*. https://media.defense.gov/2022/Jan/10/2002919460/-1/-1/1/DODIG-2022-049.REDACTED.PDF. Accessed December 14, 2024.
Verhelst, Hugo, Alexander Stannat, and Giulio Mecacci. 2020. "Machine learning against terrorism: How big data collection and analysis influences the privacy-security dilemma." *Science and Engineering Ethics* 26 (6):2975–84. https://doi.org/10.1007/s11948-020-00254-w.
Work, Robert. 2017. "Establishment of an algorithmic warfare cross-functional team (Project Maven)." Deputy Secretary of Defense. https://www.govexec.com/media/gbc/docs/pdfs_edit/establishment_of_the_awcft_project_maven.pdf. Accessed November 10, 2023.
Yee, Adrian K. 2023. "Machine learning, misinformation, and citizen science." *European Journal for Philosophy of Science* 13 (56):1–24. https://doi.org/10.1007/s13194-023-00558-1.
Zhao, Kino. 2023. "Measuring the nonexistent: Validity before measurement." *Philosophy of Science* 90: 227–44. https://doi.org/10.1017/psa.2023.3.

Cite this article: Yee, Adrian K. 2025. "Construct Validity in Automated Counterterrorism Analysis." *Philosophy of Science*. https://doi.org/10.1017/psa.2024.65