PERSPECTIVES ON TERRORISM

RESEARCH ARTICLE

More is More: Scaling up Online Extremism and Terrorism Research with Computer Vision

Stephane J. Baele,* Lewys Brace, and Elahe Naserian

Volume XIX, Issue 1
March 2025

ISSN: 2334-3745
DOI: 10.19165/2025.1559

Abstract: Scholars and practitioners investigating extremist and violent political actors' online communications face increasingly large information environments containing ever-growing amounts of data to find, collect, organise, and analyse. In this context, this article encourages terrorism and extremism analysts to use computational visual methods, mirroring for images what is now routinely done for text. Specifically, we chart how computer vision methods can be successfully applied to strengthen the study of extremist and violent political actors' online ecosystems. Deploying two such methods – unsupervised deep clustering and supervised object identification – on an illustrative case (an original corpus containing thousands of images collected from incel platforms) allows us to explain the logic of these tools, to identify their specific advantages (and limitations), and to subsequently propose a research workflow associating computational methods with the other visual analysis approaches traditionally leveraged.

* *Corresponding Author: Stephane J. Baele, University of Louvain. E-mail: stephane.baele@uclouvain.be*

Introduction

Technological advances, such as the growth of the internet and its evolution towards web 2.0, increasingly powerful computer hardware, the development of user-friendly software interfaces, and now AI-powered tools, have significantly changed our socio-political landscape. Extremism and political violence are no exception: movements from across the world and the ideological spectrum have a history of adapting to technological innovation and embracing the newest IT technologies to better organise, finance themselves, and communicate.[1] The spectacular growth of the far-right online ecosystem[2] or the deployment by ISIS of a "full-spectrum propaganda"[3] are the two most visible examples of radical political actors taking full advantage of digital tools to proselytise and recruit, build support communities, polarise target societies, and intimidate purported enemies.

This evolution is synonymous with escalating amounts of extremist content being circulated online. Berger and Morgan, for example, estimated there to be 46,000 accounts supporting ISIS on Twitter between September and December 2014, amounting to millions of messages –[4] with Twitter being only one piece of ISIS's vast propaganda jigsaw, alongside its newspaper, magazines, videos, books, songs, and many more. Likewise, Horta-Ribeiro and colleagues' exploration of the "manosphere" gathered no less than 28 million posts from more than 50 misogynistic online communities (which is only a segment of the full environment).[5] In other words, sprawling "e-extremism"[6] generates enormous quantities of (meta-)data.

The sheer scale of this content has naturally pushed scholars to use computer-assisted methods geared at the localisation, collection, and analysis of large corpora and associated datasets, gaining a type of macro-level knowledge that granular qualitative studies could never hope to reach. Increased computational power has made investigations of extremely large and multidimensional datasets technically possible. As Grimmer, Roberts, and Stewart rightly observed (for political studies in general), while "for much of its history, empirical work in the social sciences has been defined by scarcity, [...] social scientists are now in an era of data abundance",[7] a statement that (unfortunately) holds for extremism and terrorism research. Computational methods now routinely allow researchers to provide extensive network maps of extremist ecosystems,[8] evaluate users' behaviours and influence,[9] and analyse very large quantities of texts found on online spaces to identify their main themes or track ideological evolution across time.[10] Among these tools, machine learning techniques are increasingly described as constituting a particularly useful toolkit for the detection and analysis of vast amounts of online extremist data.[11]

However, what is arguably the prime component of communication in today's extremist ecosystems – visual imagery – has remained largely untouched by researchers using computational methods. This is not surprising: the visual dimension of extremist online spaces and their ecosystems has only recently attracted sustained academic consideration, and has thus received scant attention in methodological reviews and reflexions. Schuurman's 2020 review of methods in terrorism research revealed, among others, that "online content" (including non-visual material) and "media (film)" only constitute about six percent of published scholarship; the words "images" and "visual" don't even feature in the review. A recent special issue of *Studies in Conflict & Terrorism* dedicated to methods in terrorism studies did not include a contribution specifically dedicated to the study of imagery. Visual methods are similarly absent in terrorism research methods manuals,[12] and rarely feature in general political science methodology textbooks.[13]

Yet visual analyses of terrorism and extremism do exist and are gaining in popularity, already yielding important insights mostly through qualitative or "quasi-quantitative"[14] methods investigating small samples. While this is a welcome development, we argue here that, given the large empirical universe warranting analysis, much is to be gained by complementing these traditional approaches with computational analysis, doing for images what has already been done for text. ISIS, for instance, published almost 15,000 photographs and more than 1,700 photo reports between October 2014 and October 2015,[15] and this already large quantity only equates to the volume of images shared daily on certain far-right sites, forums, and social media accounts. The situation is now exacerbated by the multiplication of digital bots and AI-generated imagery.[16] The argument we make, therefore, closely parallels the one made by Grimmer and Stewart ten years ago in their seminal *Text as Data* article,[17] and extends Joo and Steinert-Threlkeld's recent *Image as Data* contribution, which transposed the former's logic to the computational study of images in political science broadly speaking.[18] As Joo and Steinert-Threlkeld argued, "images are key drivers of political phenomena, and we would do well to take advantage of new techniques to analyse them in large quantities in research".[19] In this paper, we follow and deepen this line of thought to tailor it to the field of security, terrorism and extremism studies, examining how it adapts to the specificities of extremist/violent political actors' communications. Specifically, we clarify how unsupervised and supervised visual computational methods work, explain how recent advances in some of these methods have drastically reduced the resources required for their implementation,[20] highlight the advantages and limitations of these methods when it comes to investigating large-scale visual extremist corpora, and specify the role of these tools to complement existing qualitative approaches within mixed-methods research designs.

To do so, we proceed in three main steps. First, we situate our endeavour within the relatively recent effort to study the visual component of extremist online spaces, tracing it back to the "visual turn" in International Relations, security studies, and terrorism research. Recognising the merits and strengths of this endeavour, we also reassert its shortcomings in a context marked by the hundreds of millions of images populating increasingly vast extremist ecosystems. Second, we succinctly explain in lay terms what computer vision is, distinguish between two main variants (unsupervised vs. supervised), and deploy one method from each of these approaches (unsupervised clustering vs. supervised object detection) on an original extremist visual corpus – more than 30,000 images collected from the incelosphere – in order to illustrate their logics and performance. Finally, we reflect on these empirical findings to reflect on the strengths and limitations of computational visual methods for the study of online extremism and terrorism, situating these tools within a coherent research workflow alongside non-computational approaches.

Strengths and Limitations of the "Visual Turn" in Extremism and Terrorism Studies

Extremism and terrorism scholars' recent awakening to the importance of visual imagery is inspired by a longstanding line of research in visual anthropology and sociology, and reflects similar undertakings in neighbouring fields, such as political communication[21] and IR and security studies,[22] where "the explicit study of visual politics has only recently begun to coalesce into a recognized area"[23] and ambitious research agendas are now set up to unpack the multiple implications of the "visual age". Following suit, extremism and terrorism studies are now starting to correct their biased "emphasis on understanding the words contained in the groups' messages rather than images"[24] and tendency to "focus on textual aspects and overlook the visual".[25] Taking stock of pioneering interventions, such as Bolt's *Violent Image* book,[26] Doerr's framework for studying visual mobilisation in conflict,[27] or Dauber and Winkler's edited volume

exploring *Visual Propaganda and Extremism in the Online Environment*,[28] a much-needed "visual turn" implementing a clear research agenda was called for by Maura Conway in an influential 2019 blog contribution.[29] As she aptly summarised, "the heavily visual nature of today's violent extremist online spaces makes this a real problem for us going forward however: we must do better."

The problem is, indeed, a thorny one: analysing extremist and terrorist imagery presents a range of difficulties that explain why scholarship has not taken off earlier or been more extensive. Storing (let alone sharing) images from terrorist organisations is prohibited or severely constrained in most research-intensive countries, pictures are often gruesome or shocking, the proximity of several extremist ecosystems (e.g. incel, far-right) with other fringe digital milieux (such as pornography or child abuse) means that visual datasets are often "contaminated" by problematic, unwanted, and, at times, illegal images, and, as highlighted earlier, strong visual methods are not a well-established part of the standard toolkit. Additionally, today's highly dynamic digital environments mean that images, such as memes, are not static singifiers and constantly evolve as they spread,[30] often resulting in shifting meanings or situations where the image itself being adapted in multiple ways (the prime example of this in extremism research is the "Pepe the Frog" meme, discussed below). Computational techniques have additional issues, well presented in Scrivens and colleagues' recent exposé on the challenges of online data collection in terrorism and extremism research: they involve technical skills, face opaque and ever-moving access points to platforms APIs, necessitate large storage spaces, and sometimes expensive hardware (such as servers or GPUs) and suffer from off-the-shelf commercial crawlers' inability to harvest pictures.[31]

Despite these difficulties, studies of the visual dimension of extremism and terrorism have emerged and typically belong to one of two types of what we could call *human* methods, in contrast to the *computational* methods presented below. In the first type, one or several visual tropes are hand-coded in all or (more likely) a sample of images constituting a visual corpus. This approach allows for measuring how frequently this/these tropes occur within, typically, a few hundred images (rarely beyond the thousand). In the second type, an in-depth interpretive analysis of one or a handful of particularly important, iconic image(s) is conducted to explain their symbolism, communicative function, embedded narrative, or/and emotional appeal. The first approach is more suited for studying the manifest meaning of images (e.g. the image contains a particular object or symbol), while the second is more apt at uncovering latent meaning (that is, meaning not immediately apparent but produced through the interplay of the image content and its cultural and ideological contexts and subtexts). Both types of analysis are usually deductive, in the sense that they rely on theoretical frameworks (be it from sociology, philosophy, media studies, etc.) to infer which visual tropes are *a priori* important, and subsequently look for them. However, a dose of induction is almost always present as experts make educated guesses based on previous encounters with similar datasets as to what types of images could be critical or frequently occur, sometimes at odds with theoretical expectations.

These two human methods have delivered important empirical insights when it comes to extremist and terrorist imagery. Among them, five stand out. First, each terrorist or extremist group has a specific visual style, understood as its "basic choice of the content and type of narrow visual landscape it tends to favor", and this style is sensitive to external (e.g. material constraints, political environment) and internal (e.g. leadership change, ideological evolution) shifts.[32] Baele, Boyd, and Coan's study of ISIS's visual style, for example, examined the chronological variation of four visual tropes in the organisation's imagery,[33] which has since been examined with much greater depth by Winter in a decisive volume dedicated to terrorist imagery.[34] Second, recurring images – such as those of "good Muslims" in ISIS magazines,[35] ideal citizens facing a threatening "other" in the Danish People's Party's internet,[36] or "strong virile white farmers,

traditional earthy homesteading moms" in neo-Nazi Instagram accounts–[37] play a key role in the construction of ingroup and outgroup stereotypes and "socially typified personae",[38] creating "a divide between the in-group and purported out-groups",[39] and constructing an underlying "moral order".[40] Third and more broadly, the array of visual tropes being used by extremists and terrorists is not vast; a limited number of image types is present across ideologies, each serving a particular function in outreach and subsequent radicalisation. Besides stereotypical images of ingroup and outgroup members, both symbols and gruesome/grotesque imagery –[41] including the "about to die" image[42] and more generally the "death" representation –[43] feature regularly.[44] Fourth, visual icons travel across national contexts, with processes of "translation" undertaken to make them resonate with local audiences and stakes. For instance, Doerr traced the itinerary of the "black sheep" poster from its initial Swiss context to its subsequent variations in Italy and Germany,[45] and more recently studied the recycling of US Capitol uprising images by German far-right movements.[46] This dissemination/translation process can even occur across ideological boundaries; for example, Miotto showed the similarities between jihadist and far-right visual representations of martyrdom.[47] Finally, sharing and consuming extremist imagery is, from a perspective inspired by Bourdieu and other sociologists, a social practice that participates in processes of group affiliation, self-identification, positioning, and emotional bonding, and thereby radicalisation. DeCook has, for example, shown how the display and dissemination of memes and other forms of images constitute key socialisation practices among the Proud Boys.[48]

Three key strengths of human approaches have unlocked these insights. First, as mentioned, they are usually theory-driven. This means that they are geared to producing immediately relevant findings determined by the expert's conceptual understanding and research prioritisation. Second, they are excellent at offering careful characterisations and in-depth interpretations of the specific, delineated visual landscape on which the expert zooms in. Third, they allow for multi-dimensional evaluations and interpretation of latent meaning, both in the sense of attuning to images that combine several visual tropes in sometimes subtle or even cryptic ways, and in the sense of permitting the simultaneous investigation of several visual tropes. Miotto's or Doerr's abovementioned studies provide good examples of the kind of granular, interpretive scrutiny of extremist images and visual landscapes that human approaches can offer.

Yet, as Conway suggested, we can do better. These standard human methods indeed have four important limitations. First, they are not adapted to the contemporary extremist visual environment characterised by millions of images of many sorts (pictures, memes, caricatures, image-text collages, etc.). While hand-coding can be done on small coherent samples or even full corpora (e.g. all images contained in ISIS magazines or on a given far-right blog), small-n human methods are simply incapable of scaling up and gaining a panoramic view of the visual landscape. As Bleiker puts it, "methodological issues get exponentially more difficult" in a situation characterized by "a limitless number of images".[49] Engstrom's study of images from the British National Party is a case in point:[50] having gathered over 10,000 "visual elements" from its website, he had to restrict his analytical lens to only 600 images of a certain type, thus leaving aside 9,400. Second, human methods are slow, making them inadequate for a digital context characterised by a fast turnover of images (new ones constantly appearing, established ones regularly morphing, etc.). Commercial services specialised in coding images do exist and work relatively fast (some even have in-house coders familiar with extremist environments), but relying on them for monitoring projects can become onerous and expensive, and they will always be slower than a machine. Third, while human methods have the advantage of relying on human intelligence, they also incur variability. While inter-coder reliability is usually high for unambiguous, easily identifiable items (such as the presence/absence of weapons or a particular symbol in an image), scores inevitably drop as soon as a degree of interpretation is

needed. For example, reliability scores in Baele, Boyd and Coan's abovementioned paper were almost perfect when it came to identifying pictures with symbols or gruesome elements, but much lower when it came to selecting the narratives conveyed by the pictures. As a result, some scholars only look at unambiguous tropes, avoiding potentially more insightful aspects,[51] or provide results that are hard to replicate. As Joo and Steinert-Threlkeld explain, these last two weaknesses alone explain why visual studies have traditionally been highly qualitative, focusing mostly on a few iconic images.[52] Finally, sustained close examination of extremist corpora maximises researchers' exposure to potentially harmful content (gruesome imagery, graphic violent porn, etc.). This is a major issue in our field, where researcher's wellbeing has long been ignored and exposure to this extreme content has become increasingly frequent.[53]

Additionally, we argue that some of the strengths of the human methods can also be limitations. On the one hand, the usually deductive character of these approaches means that scholars may miss crucial yet theoretically unexpected dimensions. Dominant theories or the community doxa inevitably drive our empirical attention towards particular types of images, yet in a fast-moving visual landscape, other pictorial genres might have gained prominence regardless of our pre-existing concepts. On the other hand, while zoomed-in investigations of specific visual styles and landscapes allow for granular, in-depth understandings of particular visual landscapes, they are, by design, blind to the (very) big picture. Discussing interpretive versus automated text analysis, Hart usefully compared the former with sightseeing a city on foot and the latter with viewing this city from a helicopter, both approaches are needed to gain a full understanding of the place.[54]

In sum, the strengths and weaknesses of human methods for visual analysis in extremism and terrorism studies can be represented as in Figure 1 below.

Figure 1. Strengths and weaknesses of human visual analysis methods.

Deploying Computational Visual Analysis in Extremism and Terrorism Research

For about two decades now, the shortcomings of human methods for the analysis of *textual* extremist corpora have been addressed by supplementing them with computational tools, and we suggest that recent developments in computational methods mean that the same logic should now be applied to *visual* ones. Indeed, the fast developments of machine learning methods in recent years have made computer vision faster, more accurate, and more sophisticated than ever before. In the following paragraph, we briefly define computer vision and identify its two most pertinent applications when it comes to extremism and terrorism studies (object detection and clustering), which we operationalise using original visual data from incel online spaces. We then build on this exploration to summarise, in the same way we did for human methods, its strengths and limitations when applied to extremist and terrorist content.

Computer vision

Computer vision can be defined as the "subfield of AI that enables computers and systems to process visual data, such as images and videos, and generate patterns for detecting, tracking, and classifying objects".[55] In other words, the primary goal of computer vision is to "replicate human visual abilities with computational models".[56] Pushed forward by major breakthroughs, such as Krizhevsky, Sutskever, and Hinton's "ImageNet",[57] the technology of automatically recognising patterns in visual environments has by now taken an important place in everyday life: applications include running autonomous cars' navigation systems, allowing crop monitoring by drones in large farming estates, recognising faces in surveillance systems, and detecting patterns that diagnostic specialists don't see in medical imaging. Computer vision models can be either *supervised* (presented with a training dataset wherein the programmer has labelled visuals deemed important) or *unsupervised* (in this case, the model is set up to detect patterns without guidance).

Computer vision is slowly making its way into political science and media studies, paving the way for applications in our field. For example, large corpora have been automatically mined to better understand how politicians present themselves on social media (and how these self-constructions correlate with ideological positioning[58] or convey particular emotions),[59] or how political memes circulate across social media.[60] Similar tools have very recently started to be used in extremism and terrorism research, yielding important findings. In a noteworthy contribution, Zannettou and colleagues designed a "processing pipeline" aimed at identifying and tracking the online dissemination of far-right memes, combining computer vision techniques such as clustering with other computational tools; deploying this pipeline to 160 million images from sub-Reddits, Chan boards, and Gab channels demonstrated not only the performance of the method but also the novelty of the perspectives gained through computer vision. Finkelstein and colleagues subsequently applied the same process to evaluate the frequency of the "happy merchant" meme (detected automatically in over 7 million images from a range of far-/alt-right online spaces), used as a proxy for antisemitism, and measured how this meme has gradually "infected" other visuals in memetic combinations (appearing together with the likes of Pepe the frog or Wojak in single images).[61] Using a different approach, Crawford, Keen, and Suarez-Tanguil categorised more than 135,000 images from various Chan racist boards into several types to help define their "visual cultures",[62] while O'Halloran and colleagues used qualitative multi-modal annotation to direct large-scale AI detection of ISIS-related images and text.[63] Despite their merits, these breakthrough studies remain rare and scattered, are usually published in computer science outlets, and their exact role or contribution within the field is rarely unpacked. For this reason, we try to expose as clearly as possible for the non-expert

audience the strengths and weaknesses of the two types of computer vision models holding, we argue, the most promise for extremism and terrorism studies, and on that basis we nail down their role in the field.

Data

To illustrate these methods and ground the discussion in the empirical reality, we use a corpus of more than 32,000 images featuring in posts extracted from nine different online spaces (spread across three platform types) pertaining to the so-called "incelosphere" (see Table 1 below).[64] The images were collected using a series of custom-built web-scrapers developed in the Python programming language, utilising common packages, such as *ScraPy* (https://scrapy.org/) and *Requests* (https://pypi.org/project/requests/). Depending on the platform, various APIs were used to aid data extraction where appropriate (i.e. the Telegram scraper utilised the Telethon API; see https://docs.telethon.dev/en/stable/). While definitely at the lower end of the size scale that computational methods are called to handle, this corpus possesses two advantages for the purpose of this article: firstly, the visual landscape of the incelosphere has not yet been thoroughly studied, meaning that we demonstrate our conceptual points on a novel case, and secondly, this is a hard case against which to test the computers, meaning that the corpus images appear highly diverse, cryptic, and regularly shocking.

Table 1. Corpus statistics: Number of images extracted from each incel online space.

Source of images (Online space)	Platform type	Number of images extracted
4chan/R9k	Chan board	25,747
9chan/Leftcel	Chan board	1,744
BlackPillsBasedGlobal	Telegram	1549
Blackpilled	Telegram	376
Incel	Telegram	877
IncelsCo	Telegram	624
incels	Instagram	18
B_l_a_c_k_p_i_l_l_e_d	Instagram	49
Blackpillmemez	Instagram	525
Involuntarycelibacy	Instagram	212
#Blackpill	Instagram	866

Image clustering

A first type of computer vision model perform *clustering*, which consists of "grouping together similar items into distinct clusters, so items within a single cluster are similar to each other and different from items outside the cluster."[65] In other words, clustering seeks to "partition observations into mutually exclusive categories, or clusters, using principles of unsupervised dimension reduction."[66] As such, this visual analysis technique is akin to unsupervised topic modelling in text analysis: it inductively "reveals" the major dimensions and tropes of a large corpus, allowing the researcher to "let the data speak". Four more specific similarities between text topic modelling and visual clustering illustrate the nature and relevance of the tool. First, like topic models, a range of algorithms are available when it comes to visual clustering (e.g. Affinity Propagation, Agglomerative Clustering, K-Means clustering), each implementing a different strategy for calculating visual distance/closeness and the subsequent grouping of images. Each comes with its strengths and limitations and higher/lower pertinence for particular corpora and research objectives. Second, just like the number of topics can be suggested by

the researcher conducting a text analysis, some visual clustering models require users to pre-define a given number of clusters, for the computer to subsequently calculate the optimal categorisation of images within them. Third, by assigning each image a class/category, visual clustering algorithms offer a quantification of which ones of these categories are particularly prominent in a corpus; the researcher is thus able to quickly see which types of images are (not) frequent. Fourth, although visual clustering models are unsupervised, their application to specific types of imagery can be preceded by training with a narrower visual dataset similar to the one to be examined, which increases their robustness.[67]

For these reasons, this tool is particularly useful in two strategies. First, it can help researchers gain a general overview of the visual landscape under investigation, with the main types of images appearing together in groups with measures of their respective importance. Just like a text topic model provides a quick understanding of the major themes of a corpus, a visual clustering model offers a rapid panorama of its visual landscape. This strategy is particularly useful for new and very big corpora for which little is known, i.e. where no "natural" categories are yet known. As Grimmer and Stewart explain in their work on unsupervised text topic modelling, while supervised methods assume a well-defined set of categories (inferred from existing scholarship or theoretical hypotheses), in some cases that set of categories is hard to derive beforehand.[68] Second, it could be used as a way to gain new, original knowledge of a big corpus which tends to be repeatedly appraised through the same lenses. Dominant theories (or doxa) that consistently direct empirical studies towards a limited number of visual tropes are at risk of being blind to important but unaccounted aspects, and biased in favour of particular features that may actually be found only in a minority of images. To use Grimmer and Stewart's work on unsupervised text analysis again, these "methods are valuable because they can identify organizations of text that are […] understudied or previously unknown".[69] This second strategy is, therefore, most useful when dealing with large corpora that have already been investigated by researchers using similar theoretical frameworks and associated methods, as a way to avoid bias and unlock new knowledge.

Following the successful application of this method to large visual political corpora (including Joshi and Buntain's clustering of 15,000 images posted on social media by over US politicians, which demonstrated that the types of images shared by these politicians were a "strong indicator" of their ideological position),[70] we ran a deep image clustering algorithm on our visual corpus. We used the Unsupervised Deep Embedding for Clustering (DEC) model, as originally proposed by Xie and colleagues.[71] In recent years, DEC models have proven themselves effective at clustering images into groups within large datasets. However, it has also been shown that the relationship between the models' convergence and their hyperparameters is tricky, and therefore requires multiple runs with different hyperparameter settings to achieve the optimal grouping. Crucially, there are no rules for finding the "best hyperparameter values". Instead, the researcher must ascertain this themselves through experimentation (and substantive expertise) with the number of *epochs* and *batch size* (the former refers to the number of passes made over the whole dataset, and the latter refers to the number of individual images the model sees before being updated).

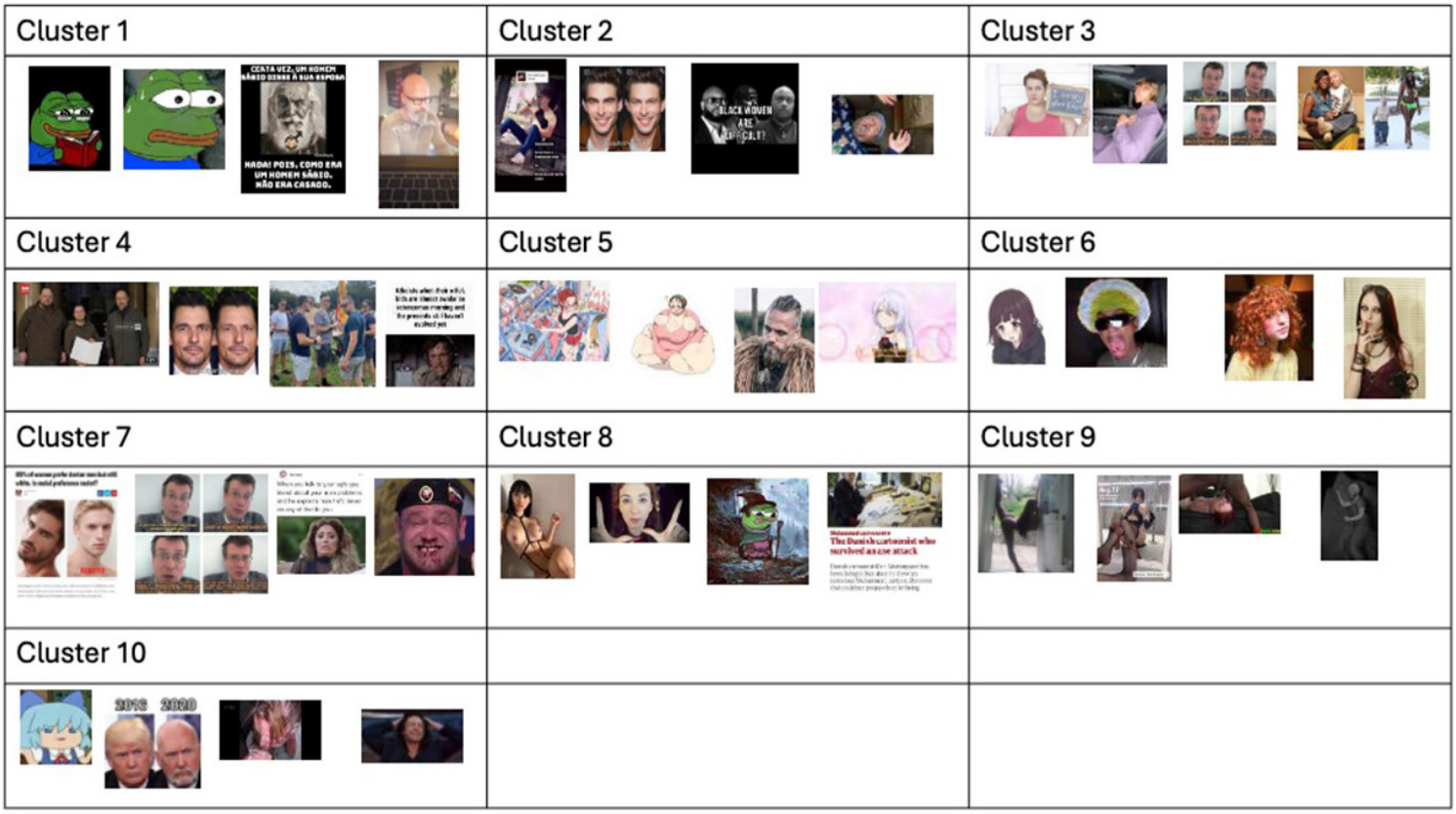
It is therefore recommended that a test set is ready to evaluate the model's output after training., as although increasing the number of epochs will tend to increase accuracy, this might be the result of overfitting (i.e. the model is fitted too closely to its training data and therefore unable to make predictions when presented with new data). Having a test set of data that the model has never seen before allows to validate the model's outputs. To demonstrate the importance of this, we trained the DEC model on the 32,587 images of our corpus for a number of different epochs and batch sizes. Two observations can be made from Table 2 below. First, the highest amount of accuracy in predictions on the test dataset that this model is able to achieve is 31

percent. Second, this is achieved using a batch size of 32 (the default batch size for most open-source model implementations), and increasing the number of epochs does not improve this. Indeed, the model only achieves 21 percent accuracy with 50 epochs at 32 batch size, but hits its highest accuracy with 100 epochs, and does not improve beyond that. We can, therefore, see that our model follows the traditional wisdom of the machine learning community, in that having more epochs does not necessarily equate to a more accurate model.[72] We can thereafter conduct a visual inspection of these results by extracting a sample of images from each cluster. Looking at Figure 2 below, we see that a model trained for 20,000 epochs with a batch size of 32 has started to create some coherence, with Pepe memes appearing a lot in cluster 1, and clusters 2 and 5 containing mostly pictures of men and anime girls, respectively.

Table 2. Comparison in accuracy at predicting data test dataset (from our incel visual corpus) for various epoch and batch size values.

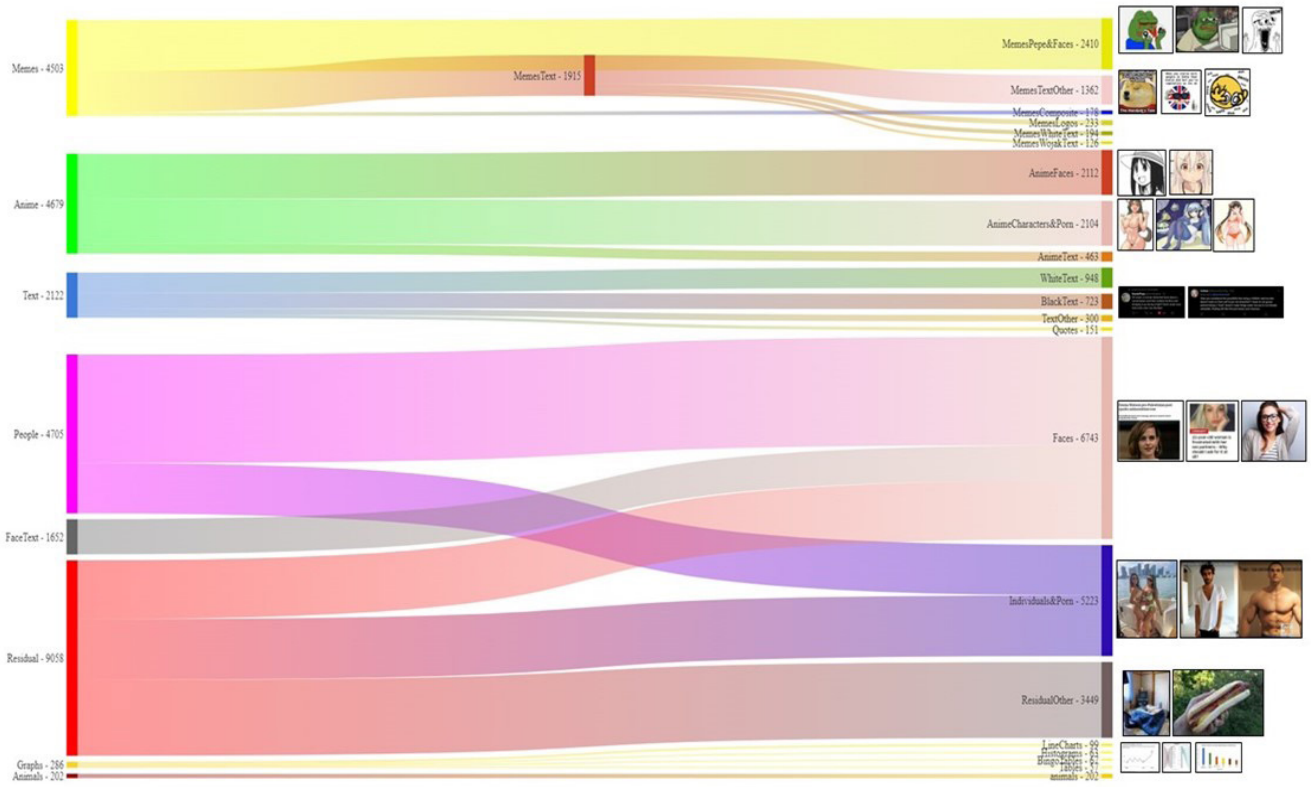
Number of epochs	Batch size	Accuracy
50	32	0.21
	100	0.05
100	32	0.31
	100	0.18
500	32	0.31
	100	0.08
2,000	32	0.31
	100	0.04
20,000	32	0.31
	100	0.03

Figure 2. Test clustering with ten clusters of incel images.



Clustering, therefore, rests on a series of trial-and-error, back-and-forth steps seeking to increase coherence, estimated not only mathematically but also through the lens of the researcher's substantive expertise. Importantly, clustering can also be re-run within clusters – for example, one could run a model within a hypothetical "vehicles" cluster to see whether the machine identifies and separates cars, planes, etc. We adopted this strategy with our whole corpus to increase coherence, and the results are presented in Figure 3 below. This is a graph inspired by Rogers' suggestion to create "metapictures", that is, ordered displays of collections of images scraped from social media in arrangements such as "tree maps" or "cluster maps" that allow for rapid scanning and critical reflection (this approach "nestles itself between qualitative visual analysis and interpretation [...] and quantitative knowledge visualization").[73] We can see that this strategy improved coherence, with the computer for instance grouping all meme characters together, as well as anime faces, anime porn characters, or graphs/ figures. Appendix 1 offers another example using a different corpus, to provide an extra illustration beyond the specific case at hand here.

Figure 3. Clustering diagram of images from the incel online ecosystem.



Four observations can be made on the basis of our application of visual clustering to the images contained in our corpus. First, the tool's ability to quickly gain a zoomed-out overview of a large visual landscape is confirmed. With no existing scholarship on incel visual imagery, our incel cluster map reveals three major types of images populating the incelosophere: people's faces (mostly women's), anime (mostly girlish characters, regularly in sexualised or pornographic), and individuals (incels, "chads" and "stacies").[74] These proportions reveal the community's valued pictorial practices (e.g. sharing erotic/pornographic and anime images, or photos of particularly attractive/repulsive individuals) and offer an indication of its major themes and ideological tenets (e.g. "lookism").

Second, the clustering tool is, as anticipated, useful for disclosing new patterns not foreseen by theory or the literature. As noted above, the scholarship on extremist visual imagery tends to

concentrate on the role of symbols and in-/out-groups depictions and, more recently, memes (which, in a sense, merge symbolism with group identification). Even though the importance of group portrayal is confirmed in our data, the cluster map also shows that cartoonish memes are not (quantitatively) a crucial visual genre, and reveals pictorial types so far ignored by the literature yet obviously empirically and theoretically relevant (e.g. "scientific" graphs, figures, and infographics).

Third, the tool's fully inductive logic inevitably produced residual clusters that are frustratingly irrelevant, theoretically speaking, with the "image-text" genre derailing parts of the process in the far-right case provided in Appendix (note that some AI models are geared at isolating images from their added texts). While theory blindness does bring about novel, unforeseen insights and observations, it comes at a cost when complex visual corpora such as the incel one are involved, which contain composite imagery for which other computational tools are required.

Fourth, the output clearly calls for additional, fine-grained research in the form of human methods (small-n hand-coding, very small-n interpretive analysis). The incel cluster map, for instance, groups together visibly different images of women serving distinct functions for the incel community and worldview that ought to be investigated.[75] As we discuss below, this method is, therefore, best used as a tool for preliminary data exploration.

Object detection

The second type of computer vision model we highlight here performs *object detection*, that is, the "task that deals with detecting instances of visual objects of a certain class (such as humans, animals, or cars) in digital images",[76] or, put differently, the "task of predicting the class and the location of different objects contained within an image."[77] The past decade has witnessed "a rapid technological evolution of object detection and its profound impact on the entire computer vision field",[78] mostly due to the application of sophisticated deep learning techniques that led to a leap forward in models' accuracy and speed,[79] including when detecting objects on moving images (videos, real-time footage). There are two types of deep-learning object detectors, which reflect developers' preferences on either end of the accuracy/speed trade-off: "two-shot" models privilege accuracy by first finding potential objects in an image and then classifying each one of them (e.g. "car", "person"), while "one-shot" models privilege speed by merging the two tasks.

Object detection is useful in two different research strategies. First, it can be used to offer a broad-brush understanding of the relative importance of a large range of standard objects in a large corpus: does this dataset contain, for example, lots of images of infrastructure/buildings or vehicles? In extremism and terrorism studies, models can quickly detect specific types of theoretically relevant images in a large visual corpus, automating (that is, speeding-up and scaling-up) what would have otherwise been a tedious hand-coding process. For example, pre-trained models automatically detect firearms in images, and models further trained on weapons are able not only to detect them but also to distinguish, in real-time video recordings, between different types of guns, rifles, knives, etc.[80] Second, models can be custom-trained to detect, and measure the prevalence of a limited number of highly specific and relevant categories of objects; i.e. Nazi symbols, ISIS magazine covers, pictures of known terrorists, particular versions of a well-known meme like Pepe the frog wearing an SS uniform, etc. Such custom-trained object detection models, which require researchers to manually code classes and their locations within images with a bounding box, are better suited to detecting specific imagery compared to unsupervised approaches, such as clustering. They are a useful tool for a range of different

study designs; for example, they can ground an evaluation of the ideological character of an online platform or serve as a measure of the ideological evolution of a given online platform or ecosystem over time (e.g. does it contain an increasing/decreasing quantity of Nazi symbols?).

To demonstrate how useful object detection models can be in studying extremist online communities, we utilised a one-shot model known as YOLO (acronym for "You Only Look Once"). Originally developed by Redmon and colleagues,[81] YOLO models depart from other object detection models that repurposed classifiers for the task of object detection, and instead treat the task as a regression problem between bounding boxes and class probabilities. This is achieved by using a single convolutional neural network (CNN) to predict both boxes by looking at the full image in one go, meaning that predictions are based on the global context of the image. YOLO is one of the most popular object detection models, due to its high level of accuracy and speed – two qualities that have made it particularly suitable for real-time applications. Given that this is a supervised learning approach, meaning that the model is presented with both training images and their associated labels as to what classes (i.e. "man", "woman", "pepe") are contained within the image and their location, the first step in developing a custom-trained object detection model is to build the training dataset. The final training dataset will consist of the images you wish to train the model on, whereby for each image there is a text file in which each line details a class and the location of the box that contains the class instance. Creating this training dataset involves manually assessing each image and creating a corresponding bounding box for any of the classes that the researcher wishes their model to detect, with these bounding boxes being where the class appears in the image.[82] Figure 4 below provides an example of what this process looks like for four different images. This time-intensive process of labelling classes in training images means that, despite the significant impact supervised object detection models could have for extremism researchers, there are some aspects that researchers with limited resources might struggle with.[83] However, new models still get increasingly powerful, continuously reducing the time and expertise needed for training them. In our case, YOLO proved powerful enough to learn specific theory-relevant objects on a relatively small training dataset.

Figure 4. An example of manually created bounding boxes for three different classes in four images.

Specifically, we followed a two-step training process using 11,700 images randomly selected from the whole corpus, in order to ascertain what our classes of interest would be. First, 4,000 images from the random sample of 11,700 were (again randomly) selected and evaluated for potential classes of analytical interest (i.e. classes that appeared regularly enough and/or

contained some known specific meaning within the incelosphere), which was done in order to control for the fact that many of the incel images were random in nature and did not have any real significant meaning or content to them. The results of this step were then evaluated, with the frequency of each class being assessed and removed if it was below ten;[84] this resulted in twelve classes. In a second step, we took these twelve classes and labelled any instances of them that appeared in the 11,700 sample by creating a bounding box as detailed above. This resulted in a total of 3,620 labels across 2,808 images (detailed in Table 3 below); 8,892 images contained miscellaneous content that did not fit into any of the twelve classes of analytical interest. Once complete, the labelled images were divided into 2,307 training and 501 test images.

Table 3. Frequency of instances per class in the incel training dataset.

Image category label	Number of ground truth boxes
Man	1066
Anime girl	783
Woman	656
Social media post	324
Pepe	270
Nude woman	146
Wojak	115
Porn	111
Military	61
Chad	60
Trad girl	15
Swastika	13

An instance of Ultralytics' YOLOv8[85] was then implemented and trained on the sample for 500 epochs, yielding impressive results given the small dataset it was trained on (see Appendix 2 for additional information on validation metrics). Figure 5 below shows a precision-recall curve graph, which plots the precision and recall in our model for each of the classes for all possible user-defined thresholds between 0-1. Ideally, precision and recall would be high (in other words, the model detects all instances of a category, and what it detects is correct). However, in reality, there is a trade-off between precision and recall since a lower threshold decreases the chances of ground truth boxes being missed since this would increase the number of detections made by the model, but at the same time, a higher threshold would mean that the model is more confident in what it predicts. Thus, a precision-recall curve is useful in ascertaining what the user-defined threshold should be, depending on the desired application of the model. In Figure 5, we show the conditional probability that the bounding box belongs to the specific class it is next to. For the purposes of this paper, we will also use a probability of 0.5 to signify a hypothetical baseline model; in other words, like logistic regression, if an instance is believed to belong to a specific class with a probability of >0.5, it is tagged as positive. This baseline model is represented by the grey dashed lines. Figure 6 below additionally supports what we see in Figure 5. Figures 7 a, b, and c below, which are collages of predictions our trained model made on the 501 test images (i.e. images that the model did not see during training), also allow for manual inspection of the model's accuracy.[86]

Figure 5. Precision-recall curve for the custom-trained YOLOv8 model trained on the incel dataset.

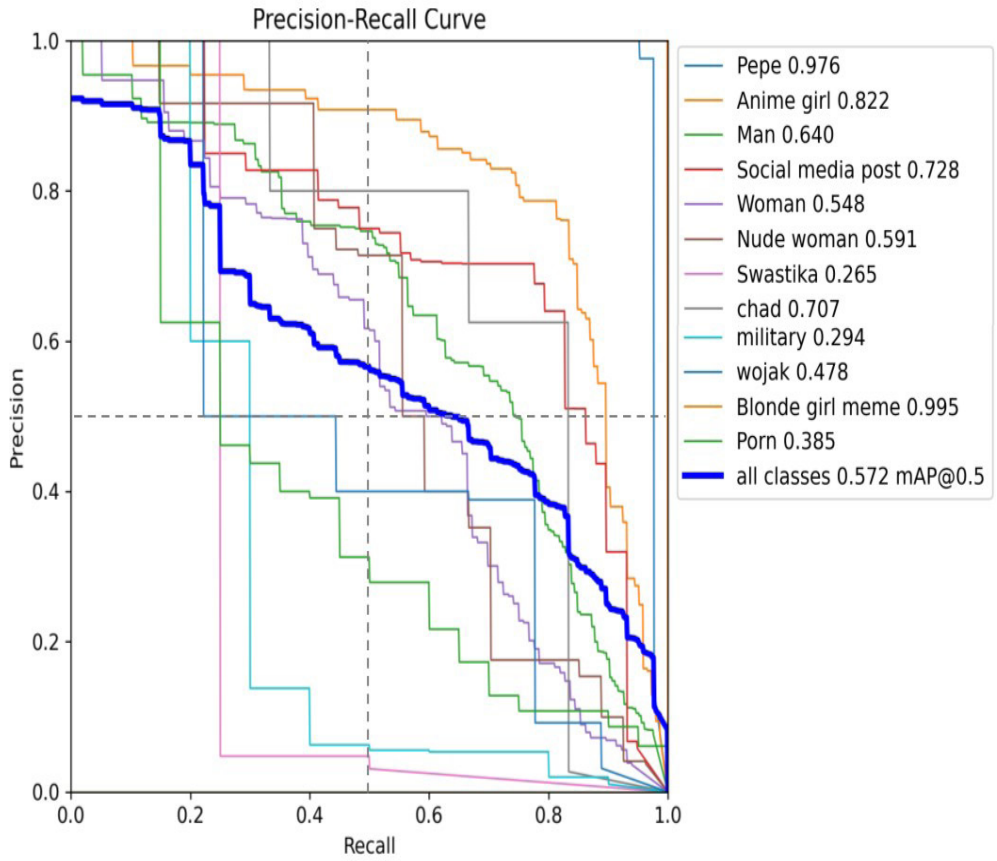
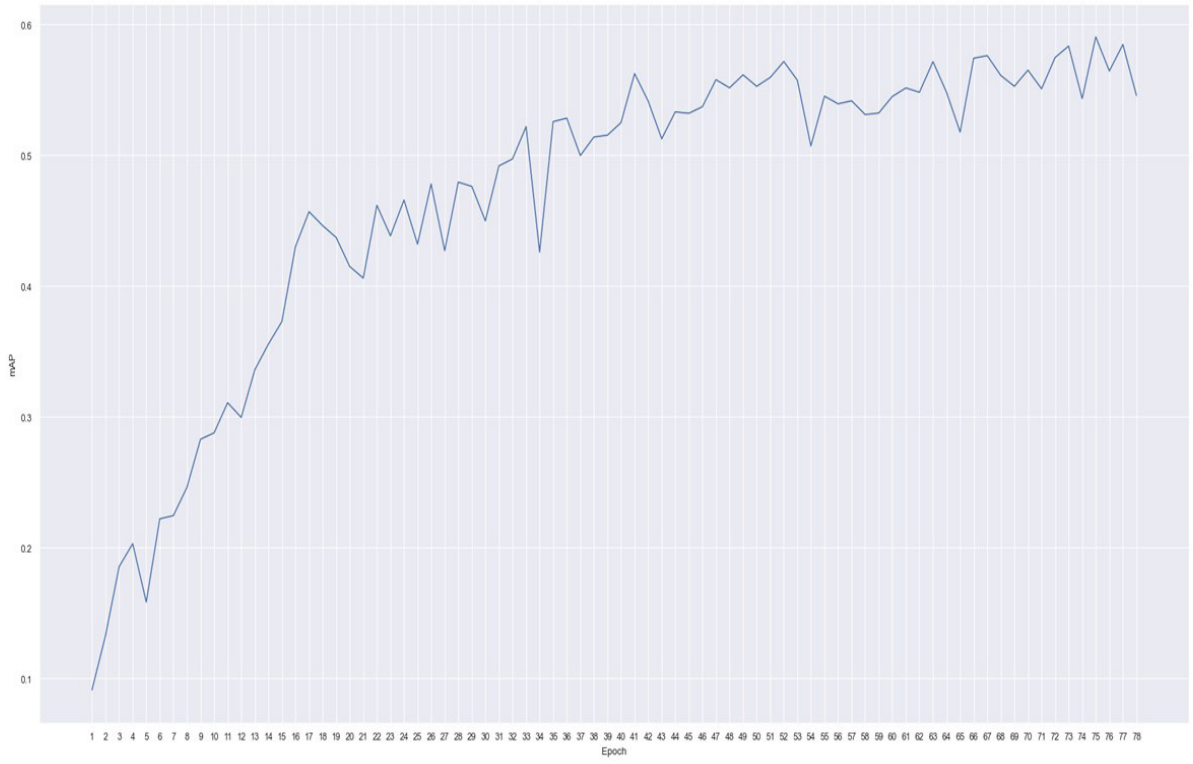


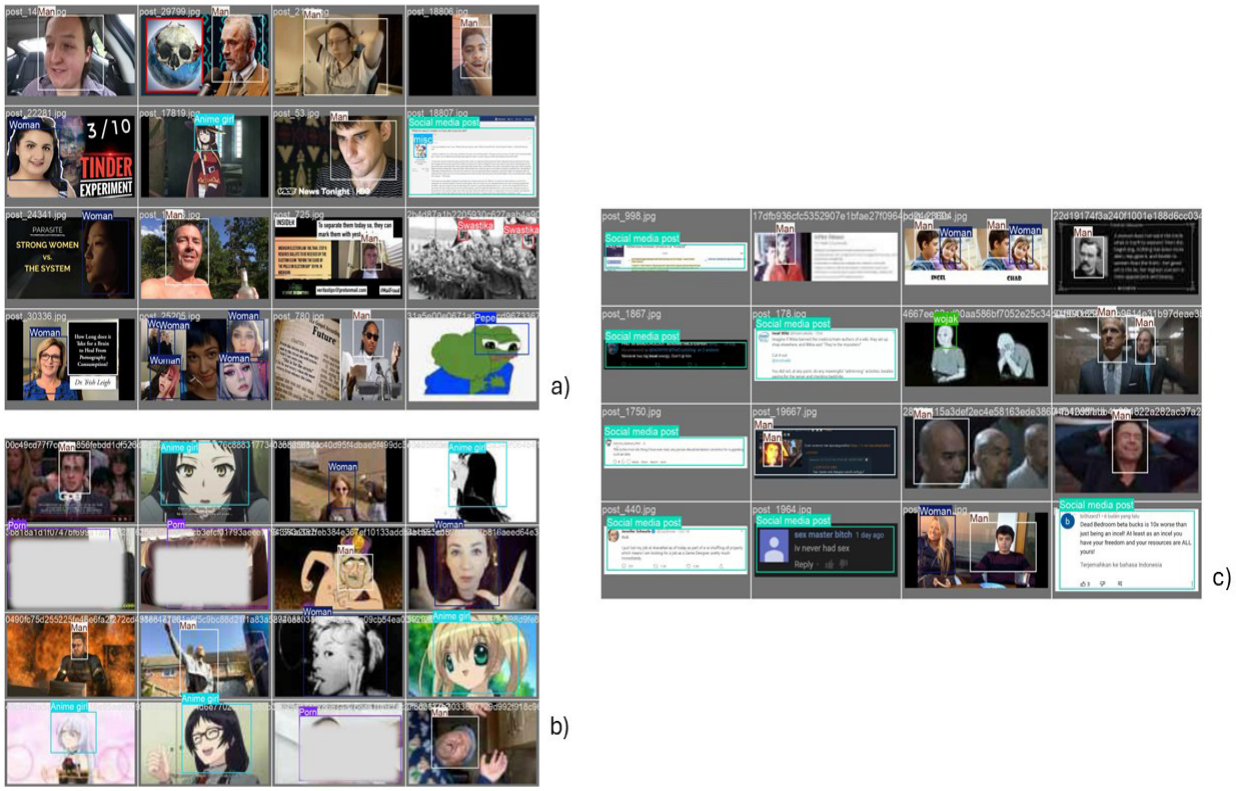
Figure 6. mAP score for the custom-trained YOLOv8 model across epochs (1-75).



Looking at Figure 5, we can see that the model has developed near-perfect precision recall for both the "Pepe" (0.976) and "Trad girl" (0.995) memes (their respective lines almost follow the exact shape of the top right-hand corner of the graph box). The high confidence value for these two memes shows the importance of thinking about bounding boxes when undertaking an object detection project looking at memes. Namely, as stated above, within the incel (and many other) online subcultures, the same memes often appear in different forms. This is particularly true with the Pepe meme, which is not only shared in its "base form" but is also wearing hats, having facial hair, sometimes having a whole body, and many other forms. This can make it difficult for this image to be detected by computer vision algorithms. However, due to the use of bounding boxes in object detection, we were able to specify that the model should look for a specific part of a meme to detect it — in this case, the bounding boxes were set to the Pepe character's eyes as these were, in a sense, the smaller common denominator in the vast majority of Pepe meme iterations (cf. Figure 4 above). As can be seen in Figure 7 a below, this enabled the model to accurately detect this meme, despite its numerous forms and only 270 coded instances of it in the training dataset. Impressively, the model's success in finding Trad Girl memes rested on only 15 coded instances in the training dataset; here the model's power was aided by the fact that this meme has unique visual features and appears with little variation in the dataset.

The model also proved to be quite effective at detecting other classes. With 783 instances in the training dataset, the model developed a 0.822 confidence score for the "anime girl" class, likely due to the highly stylised nature of these images (i.e. large eyes and pointy hair). With a confidence score of 0.728 based on 324 training instances, the model detected social media posts well; it also proved relatively successful at detecting headshots of both men (0.640 confidence with 1066 training instances) and women (0.548 with 656 training instances). As can be seen in Figures 7 a, b, and c, unlike the three classes above, these classes are characterised by high variability (ethnicity, facial hair, head hair, glasses, etc) without any stylised and uniform characteristics, and would, therefore, require more training instances in order to improve the confidence score. Further, the model was able to achieve 0.591 confidence with the nude woman class based on a very small set of 146 training images, a testament to the YOLO approach given the large amount of variation between different training instances in terms of poses, etc. Likewise, a confidence level of 0.707 with only 60 training instances for the "Chad" meme is impressive. The rest of the classes fall below the 0.5 confidence of our hypothetical baseline model, but Figures 7 a, b, and c nonetheless show that some predictions are accurate. In these cases, more training instances are required to allow the model to handle variations within a single object deemed relevant.

Figures 7 a, b, c. Collages of validation images showing some of the predictions made by the custom YOLOv8 model.



Capitalising on Strengths, Acknowledging Weaknesses: Towards a Research Workflow Integrating Computer Vision

Several strengths and weaknesses of computational visual analysis clearly emerge from the application and results above, summarised in Figure 8 below.

Figure 8. Strengths and weaknesses of computational visual analysis methods.

In terms of strengths, these methods are first and foremost capable of scaling-up the analysis to process huge visual corpora containing tens or hundreds of thousands of images, which is a critical asset in today's very large digital extremist environments. The bigger the corpus, the more useful these methods are, allowing the researcher to move from small-n samples to full corpus analysis, and to do so speedily, outpacing any human coder when it comes to detecting specific objects. In our case, we were able to rapidly detect images featuring specific classes of objects (i.e. Pepe, anime girls) from a large corpus. Importantly, these methods leverage transparent and replicable procedures invariably producing similar results on a given corpus (something never guaranteed with human methods). Finally, and not insignificantly, these automated methods also limits the researcher's exposure to potentially harmful material, which responds to the concerns recently voiced about one of the key challenges when handling terrorist and extremist images (cf. above).

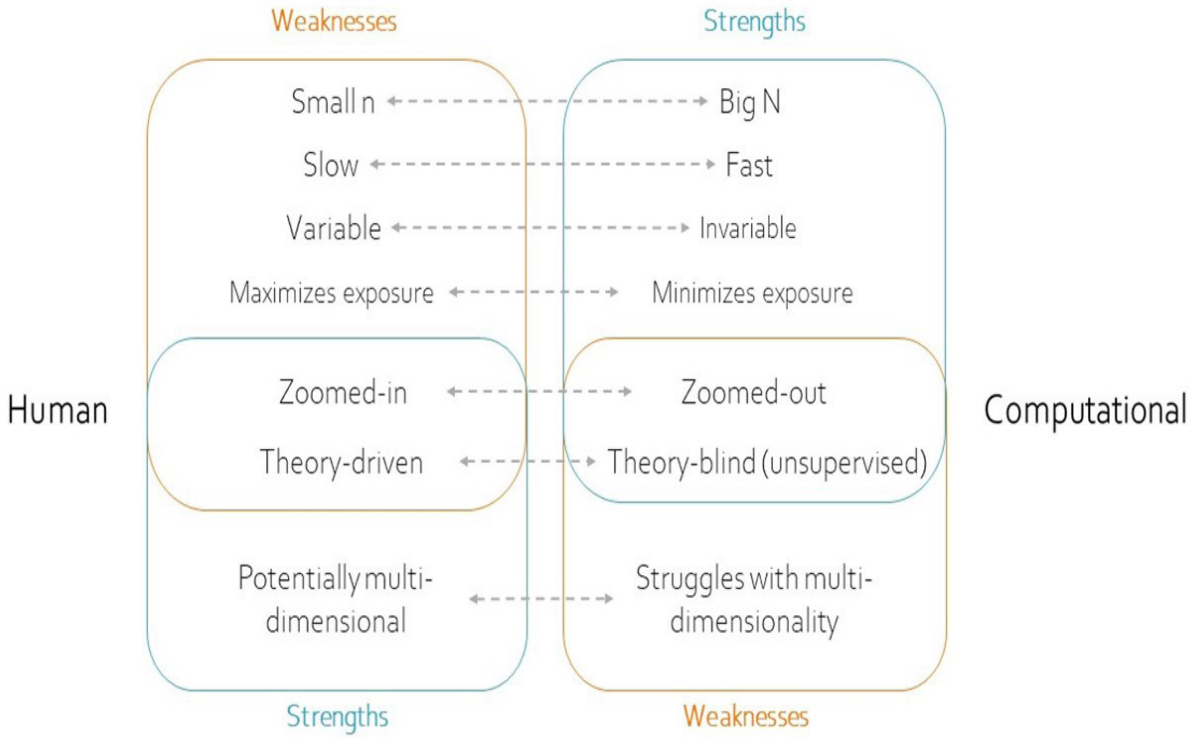
Yet computational methods also have clear weaknesses. First, even though object detection models can identify several objects within the same image, computational methods still struggle with multi-dimensionality and latent meaning, lacking the human appreciation of how certain visual signifiers blend to create new composite signs, and still needing training to identify general visual classes expressed in many different forms (e.g. a "Nazi symbol"). Second, these methods are incapable of offering detailed qualitative interpretations of particularly important images or groups of images, as they only offer zoomed-out views of vast visual landscapes. Our clustering map, for instance, revealed the presence of large numbers of anime, and even grouped them into two distinct classes, but fell short of providing any real insight into what sub-variants may exist and, ultimately, what these images mean. These first two weaknesses are significant: In spite of their strengths, computational methods simply cannot do the type of in-depth work done in some of the aforementioned studies.[87] Third, the unsupervised models are theory-blind, which means that their purely inductive logic produces results that at times lack any usefulness and contain large residual categories – when they are not outwardly counter-productive. Clustering models sometimes need to be run several times with varying hyperparameters, both on the full corpus and within clusters, before they succeed in separating pictures into meaningful categories.

Yet the last two weaknesses also constitute strengths. First, the ability to offer zoomed-out perspectives on today's sprawling extremist digital ecosystems without pre-existing bias is critical, allowing the researcher to subsequently focus their efforts on the segments of the visual landscape that appear to dominate or matter in particular ways. Second, unsupervised methods' blindness to theory at times unlocks unexpected knowledge. As Grimmer, Roberts, and Stewart explain, hypothetic-deductive approaches force "researchers to use data to test theories that were developed before the data arrive", which "leaves potential insights on the table". By using inductive computational tools as a preliminary step, we "often discover new directions, questions, and measures", a step they call "quantitative discovery".[88]

The reader has by now realised that the strengths and weaknesses of human versus computational methods mirror one another. Figure 9 below connects Figures 1 and 8 in a single graph, showing how one approach's strengths reflect the other's weaknesses. Such a conceptualisation naturally calls for human-computational mixed methods in extremism and terrorism studies, for the potential of the fields' "visual turn" to be fully realised. This proposition provides an operational response to Rodermond and Weerman's recent observation that in terrorism and extremism studies "each method has its own strengths and shortcomings, and each method has its particular problems when applied in practice."[89] Furthermore, we hope that our illustrative incel case-study has shown that in spite of its underlying technicality, productive computer vision necessarily involves "qualitative" fine-tuning, mobilising expert substantive

knowledge – in that sense, the boundary between human and computational approaches is not an impermeable one.

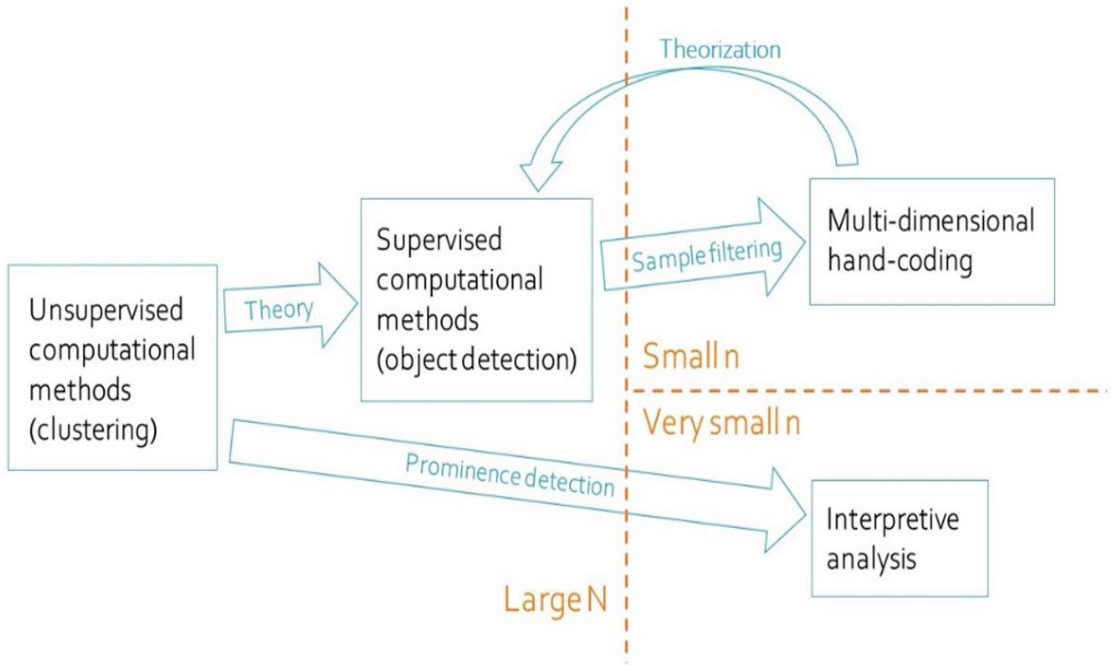
Figure 9. Strengths and weaknesses of human vs. computational visual analysis methods.



We therefore propose to interlock both approaches into a single, coherent research workflow for the study of large extremist visual landscapes, as summarised in our final Figure 10 below. As Grimmer and Stewart rightly argued, "rather than replace humans, computers amplify human abilities," meaning that a dichotomous, quantitative-qualitative (human-computational) methodological opposition is misleading.[90] Our workflow starts by running a clustering model with different parameters, allowing us to explore the corpus without bias and gain a bird's eye perspective. This zoomed-out view can direct towards two scenarios, either allowing for the direct detection of prominent or important visual tropes to then be interpreted qualitatively or to help identify or refine relevant theory to calibrate an object detection model. To take our incel case, the clustering model revealed large clusters of human faces and bodies, which could serve as a basis to train an object detection model seeking to differentiate between particular, theoretically relevant types (e.g., types of men such as "chads" and "incels", or pictures featuring violence against women). These classes of images can thereafter serve as samples for a classic multi-dimensional hand-coding of a smaller sample (e.g., a codebook-based analysis of pictures featuring violence against women), whose results could, in turn, serve to re-calibrate a computer vision model to find more theoretically relevant sub-classes. As such, our workflow advocates the same logic articulated by Grimmer, Roberts, and Stewart in their reflection on computational text analysis methods: it deconstructs and revisits the deductive-inductive and quantitative-qualitative oppositions that too often paralyse progress in social sciences,[91] adopting a dynamic stream where iterative loops ensure that data, observation, expertise, and theory enrich each other. As Bucy and Joo note, fostering such a dialogue across deeply-entrenched traditions and

academic training lines is a challenge,[92] but examples of good practice are starting to emerge; Crawford, Keen, and Suarez-Tangil, for instance, ran a clustering model on tens of thousands of images from various Chan image-boards to enrich theory and hone in a qualitative analysis of violent far-right memes.

Figure 10. A research workflow for visual research in extremism and terrorism studies.



Conclusion and Discussion: Towards Mixed Human-Computer Research Streams

The proliferation of online extremist content and digital terrorism-related communications has encouraged researchers to turn to social data science, and advances in artificial intelligence have further bolstered their use of computational methods. However, this effort has, so far, largely stayed within the confines of text and meta-data analysis, failing to unlock new horizons for the field's recent "visual turn." The present paper sought to affect this connection between computational and visual analysis by detailing how supervised and unsupervised computer vision models can enrich the study of extremist and terrorist imagery in decisive ways. We applied two of these models to an original corpus to explain and illustrate the strengths and weaknesses of computational visual analysis applied to extremism and terrorism research; because these benefits and limitations mirror those of human visual methods. We proposed a research workflow combining the two types of approaches and undermining the sterile qualitative-quantitative opposition too often preventing security studies to meet its potential.

Research on terrorist and extremist communications has greatly benefited from computational text analysis, which has effectively complemented qualitative methodologies; there is no reason why our "visual turn" shouldn't follow track. It is worth stressing, however, that this workflow doesn't exhaust the range of research questions and, therefore, methodological approaches and research agendas centring around extremist and terrorist imagery; as Bouko rightly

emphasised, the plurality of visual dynamics constituting political reality call for an array of research questions and associated methods, some of them not even interested in the content of images (focusing, for instance, on impacts).[93]

It is hoped that our workflow and underpinning discussion will contribute to further materialise the recent optimistic assessments of our field's growing methodological strength,[94] which contrasts with the older, well-known negative and dire diagnoses of poor rigour, "failings", and "stagnation."[95] Specifically, adding computer vision to our methodological toolkit has the potential to partly correct terrorism studies' overreliance on qualitative designs (which make up more than three quarters of published research) and to move scholarship standards towards a more thorough and systematic exposition and justification of their methods (which is not done in most published articles).[96]

An integrated mixed-methods designs making the most of human and computational visual methods while mitigating each one's shortcomings not only makes scientific sense, it can also serve a positive social purpose at a time when visual AI models start to be leveraged by extremist and terrorist actors themselves (as well as by governments, in a dialectical power struggle).[97] As Bucy and Joo rightly suggest, "understanding the prevalence of these visual forms of hate, and accurately assessing their presence [...] could put visual politics to a prosocial use."[98] Beyond academia, law enforcement and security practitioners have started to leverage computational techniques to respond to "the urgent need to collect and analyze terrorist and extremist content online, on a large scale",[99] but often lack the tools to do so when it comes to imagery. Image clustering models can be run to rapidly situate the ideology and culture of an online space that has recently been brought to attention (for example, as part of an investigation or after an attack), while object detection models can be trained to look for specific details such as weapons (weapons visual datasets already exist that can feed the training), ideological markers such as svastikas, faces of individuals such as extremist leaders or potential attack targets (e.g. political figures), or high-risk content such as bomb-making blueprints. However, given the high stakes of this type of monitoring and analysis, law-enforcement and intelligence practitioners should be careful about the multiple errors, risks, and biases paving the approach, collection, and analysis of big data, which are by now well-documented.[100]

Finally, while primarily methodological, the contribution of this paper is also empirical. Using a real corpus as an illustrative example has incidentally led us to offer novel observations on the visual style of the incelosphere – which had, to our knowledge and surprise, not yet been systematically and directly studied.[101] Paving the way for a more comprehensive and in-depth study of incel imagery, we can already highlight the following three observations. First, the incel visual style appears to be highly composite and diverse, quite far from the typically narrow extremist landscapes featuring ingroup and outgroup archetypes (cf. above); a lot of the dataset is made of images that cannot easily be categorised into known theoretical categories. Second, most of the images feature specific individuals (actors, popular internet personalities, politicians, etc.) who have no obvious connection to the incel community; further research would need to adapt their methods to make sense of this fact. Finally, our dataset (unsurprisingly) contains a significant number of erotic or pornographic pictures and anime. Further research could examine the gender and sexual meanings of these images, which, among others, appear to ridicule women and picture them in submissive situations (in pictures) while simultaneously construct an idealised "girlfriend" stereotype (in anime), a paradox already highlighted in the text-based literature.[102]

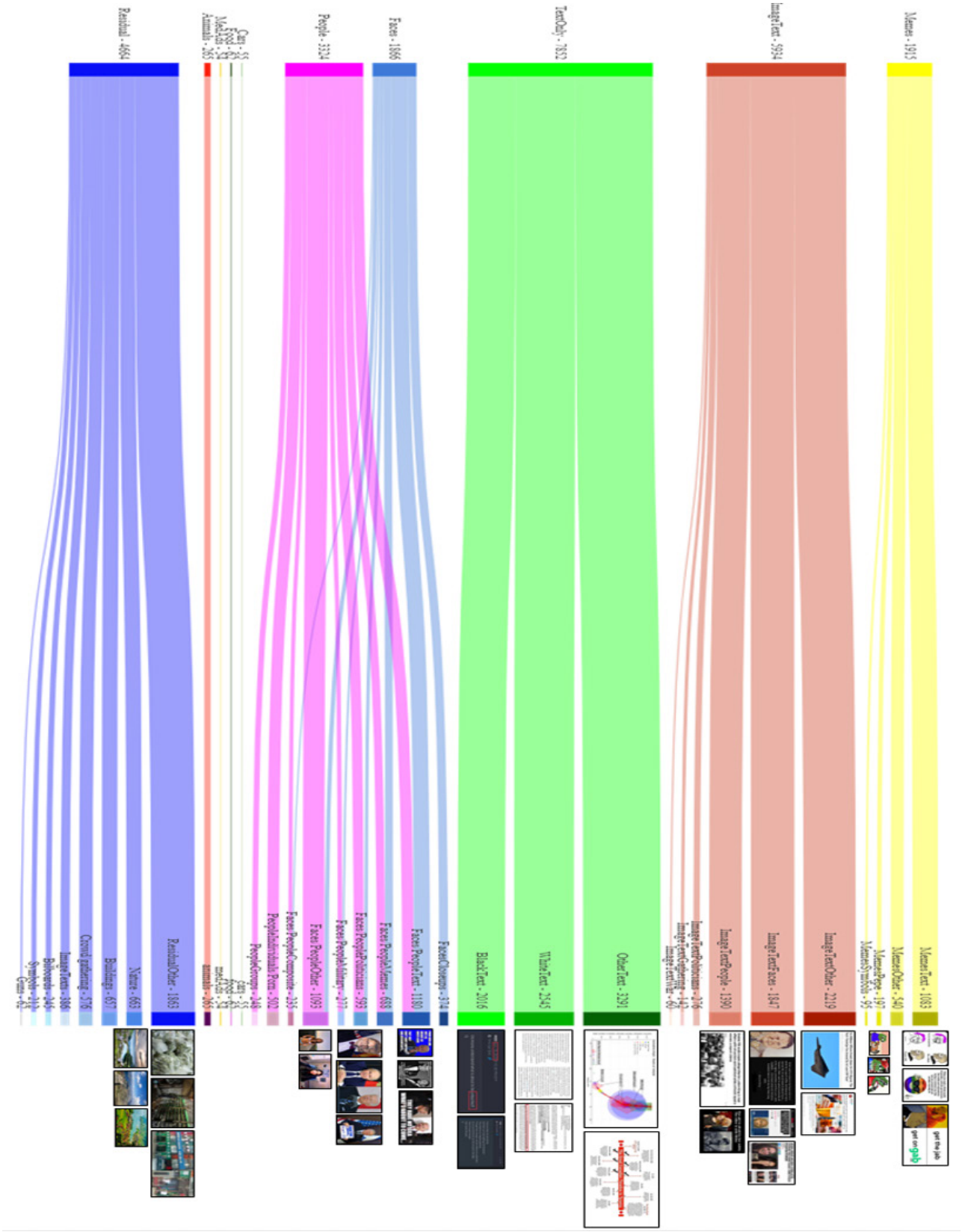
Stephane J. Baele is Professor of International Relations at the University of Louvain (Belgium), and Honorary Associate Professor of Security and Political Violence at the University of Exeter (UK). His multidisciplinary work focuses on extremist and violent political actors' communications. On Twitter at @StephaneBaele.

Lewys Brace is Senior Lecturer in Computational Social Science at the University of Exeter (UK), where he is the co-Director of the Centre for Computational Social Sciences (C2S2). His work leverages data science and OSINT methods to study phenomena pertaining to terrorism, radicalisation, and cybercrime. On Twitter at @Lew_Brace.

Elahe Naserian is a senior AI engineer and data scientist at Trilateral Research UK, which she joined after a PhD at the University of West of Scotland and postdoctoral fellowships at the University of Exeter. She specialises in the development and application of Artificial Intelligence to analyse social issues.

Appendix 1: Far-Right Telegraph Clustering

Figure A-1. Clustering diagram of images from the US Far-right Telegram ecosystem.



Appendix 2: Validation Metrics

Many of the metrics used to assess object detection models require understanding the difference between the "ground truth bounding box", the bounding box that indicates the location of a specific class instance that was labelled by researchers, and the "predicted bounding box", which is the bounding box that the algorithm believes contains a class instance. The "Intersection over Union" (IoU) is then a quantitive measure of the overlap between the two and can range between 0 (no overlap) and 1 (perfect overlap), with a higher IoU indicating more accurate object detection.

Each detection instance can then have one of three outcomes. First, it can be "true positive", meaning the model correctly identified both the category and location of the object in an image and the IoU score is equal to or greater than the specific threshold. Second, the instance can be "false positive", whereby the model incorrectly identifies an object that does not exist in the ground truth box or where the predicted bounding box has an IoU lower than the specific threshold. Finally, "false negative" refers to instances whereby the model failed to detect an object that is specific by the ground truth box.

One of the most important metrics when validating an object detection model is then the average precision (AP) score. Here, "precision" refers to how precise the model is, and it is essentially the proportion between the number of true positives and the number of predictions. Linked to this is the notion of "recall", which is the ratio between the true positives the model makes and the number of ground truths; in other words, how many instances of an object in an image the model detects. More formally:

$$Precision = \frac{true\ positives}{true\ positives + false\ positives}$$

$$Recall = \frac{true\ positives}{true\ positives + false\ negatives}$$

Endnotes

1 On extremists and terrorists' technological innovation, read for example Adam Dolnik, *Understanding Terrorist Innovation. Technology, Tactics and Global Trends* (New York: Routledge, 2007); Yannick Veilleux-Lepage, *How Terror Evolves. The Emergence and Spread of Terrorist Techniques* (Lanham: Rowman & Littlefield, 2020).

2 For a historical account, read Maura Conway, Ryan Scrivens, and Logan Macnair, "Right-Wing Extremists' Persistent Online Presence: History and Contemporary Trends," *ICCT Policy Briefs* (2019). For a panorama and theorization of the contemporary ecosystem, read Stephane Baele, Lewys Brace, and Travis Coan, "Uncovering the Far-Right Online Ecosystem: An Analytical Framework and Research Agenda," *Studies in Conflict & Terrorism* 46, no.9 (2020): 1599-1623.

3 For a comprehensive account, see Stephane Baele, Katharine Boyd, and Travis Coan (eds.), *ISIS Propaganda. A Full-Spectrum Extremist Message* (New York: Oxford University Press, 2020).

4 J.M. Berger and Jonathon Morgan, "The ISIS Twitter Census. Defining and Describing the Population of ISIS Supporters on Twitter," *Brookings Project on U.S. Relations with the Islamic World Analysis Papers* 20 (2015). See also Adam Badawy and Emilio Ferrara, "The rise of Jihadist Propaganda on Social Networks," *Journal of Computational Social Science* 1, no.2 (2018): 453-470; or Jytte Klausen, "Tweeting the Jihad: Social Media Networks of Western Foreign Fighters in Syria and Iraq," *Studies in Conflict & Terrorism* 38, no.1 (2015): 1-22.

5 Manoel Horta Ribeiro, Jeremy Blackburn, Barry Bradlyn, Emiliano De Cristofaro, Gianluca Stringhini, Summer Long, Stephanie Greenberg, and Savvas Zannettou, "The Evolution of the Manosphere Across the Web," *Proceedings of the International AAAI Conference on Web and Social Media* 15, no.1 (2021): 196-207.

6 Xinyi Zhang, and Mark Davis, "E-extremism: A conceptual framework for studying the online far right," *New Media & Society* 26, no.5 (2022).

7 Justin Grimmer, Margaret Roberts, and Brandon Stewart, "Machine Learning for Social Sciences: An Agnostic Approach," *Annual Review of Political Science* 24 (2021): 395-396.

8 See for example Manuela Caiani and Claudius Wagemann, "Online Networks of the Italian and German Extreme Right," *Information, Communication & Society* 12, no.1 (2009): 66-109; Caterina Froio and Bharat Ganesh, "The Transnationalisation of Far Right Discourse on Twitter," *European Societies* 21, no.4 (2019): 513-539; Ofra Klein and Jasper Muis, "Online Discontent: Comparing Western European Far-right Groups on Facebook," *European Societies* 21, no.4 (2019): 540-562; Aleksandra Urman and Stefan Katz, "What They Do in the Shadows: Examining the Far-right Networks on Telegram," *Information, Communication & Society* (2020), online before print.

9 For example Ryan Scrivens, "Exploring Radical Right-Wing Posting Behaviors Online," *Deviant Behavior* 42, no.11 (2020): 1470-1484; Stephane Baele, Lewys Brace, Travis Coan, and Elahe Naserian, "Super- (and hyper-) posters on extremist forums," *Journal of Policing, Intelligence & Counter Terrorism* 18, no.3 (2023): 243-281.

10 See for example Sylvia Jaki, Tom De Smedt, Maja Gwóźdź, Rudresh Panchal, Alexander Rossa, and Guy De Pauw, "Online Hatred of Women in the Incels.me Forum. Linguistic Analysis and Automatic Detection," *Journal of Language Aggression and Conflict* 7, no.2 (2019): 240-268; Ryan Scrivens, Garth Davies, and Richard Frank, "Measuring the Evolution of Radical Right-Wing Posting Behaviors Online," *Deviant Behavior* 41, no.2 (2020): 216-232; or Stephane Baele, Lewys Brace, and Debbie Ging, "A Diachronic Cross-Platforms Analysis of Violent Extremist Language in the Incel Online Ecosystem," *Terrorism & Political Violence* (2023), online before print.

11 Stuart Macdonald, Elizabeth Pearson, Ryan Scrivens, and Joe Whittaker, "Using online data in terrorism research," In *A Research Agenda for Terrorism Studies*, edited by Lara Frumkin, John Morrison, and Andrew Silke (Edward Elgar, 2023): 145-158. Miriam Fernandez and Harith Alani, "Artificial Intelligence and Online Extremism. Challenges and Opportunities," In *Predictive Policing and Artificial Intelligence*, edited by John McDaniel and Ken Pease (London: Routledge, 2021).

12 To our knowledge, there is no comprehensive, specific handbook of methods in terrorism research, but the broader terrorism studies handbooks that include methods sections do not feature visual approaches – see for example Erica Chenoweth, Richard English, Andreas Gofas, and Stathis Kalyvas, *The Oxford Handbook of Terrorism* (Oxford: Oxford University Press, 2019). Dixit and Stump's methods handbook stands out, but it solely includes "critical" approaches and its last chapter on visuals consists in an interview focusing only on visual representations of war on terror – see Priya Dixit and Jacob Stump, *Critical Methods in Terrorism Studies* (London: Routledge, 2016).

13 See for instance David McNabb, *Research Methods for Political Science. Quantitative and Qualitative Methods. 2nd Edition* (New York: Routledge, 2015).

14 By that we mean approaches that rely on the "hand" coding of limited samples of pictures (typically

between 100 and 1,000) by experts, and descriptive statistics built from this coding.

15 Stephane Baele and Charlie Winter, "From Music to Books, from Pictures to Numbers: The Forgotten Yet Crucial Components of Islamic State's Propaganda," In *ISIS Propaganda: A Full-spectrum Extremist Message*, edited by Stephane Baele, Katharine Boyd, and Travis Coan (New York: Oxford University Press, 2019): 188-218.

16 Stephane Baele and Lewys Brace, "AI Extremism. Technology, Tactics, Actors," *VoxPol Reports* (2024), https://voxpol.eu/file/ai-extremism-technology-tactics-actors/

17 Justin Grimmer and Brandon Stewart, "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts," *Political Analysis* 21, no.3 (2013): 267-297.

18 Jungseock Joo and Zachary Steinert-Threlkeld, "Image as Data: Automated Content Analysis for Visual Presentations of Political Actors and Events," *Computational Communication Research* 4, no.1 (2022): 11-67.

19 Joo and Steinert-Threlkeld, "Image as Data," 35.

20 One important reason explaining the delayed adoption of computational visual methods within extremism research is indeed the lack of adequate resources available to researchers; using some of the supervised methods discussed below, for instance, has traditionally required lots of time dedicated to labelling the dataset for training.

21 See for instance Dan Schill, "The Visual Image and the Political Image: A Review of Visual Communication Research in the Field of Political Communication," *Review of Communication* 12, no.2 (2012): 118-142; Anastasia Veneti, Daniel Jackson, and Darren G. Lilleker (eds.), *Visual Political Communication* (Cham: Palgrave Macmillan – Springer, 2019)

22 For example Roland Bleiker, "The Aesthetic Turn in International Political Theory," *Millennium: Journal of International Studies* 30, no.3 (2001): 509-533; Roland Bleiker (ed.), *Visual Global Politics* (New York: Routledge, 2018); or Lene Hansen, "Theorizing the Image for Security Studies: Visual Securitization and the Muhammad Cartoon Crisis," *European Journal of International Relations* 17, no.1 (2011): 51-74.

23 Erik Bucy and Jungseock Joo, "Editors' Introduction: Visual Politics, Grand Collaborative Programs, and the Opportunity to Think Big," *The International Journal of Press/Politics* 26, no.1 (2021): 5-21.

24 Winkler and Dauber, *Visual Propaganda*, 13

25 Robin Engstrom, "The Online Visual Group Formation of the Far Right: A Cognitive-Historical Case Study of the British National Party," *Public Journal of Semiotics* 6, no.1 (2014): 1-21.

26 Neville Bolt, *The Violent Image: Insurgent Propaganda and the New Revolutionaries* (New York: Columbia University Press, 2012).

27 Nicole Doerr, Alice Mattoni, and Simon Teune, "Toward a Visual Analysis of Social Movements, Conflict, and Political Mobilization," *Research in Social Movements, Conflicts and Change* 35 (2013).

28 Caroline Winkler and Cori Dauber (eds.), *Visual Propaganda and Extremism in the Online Environment* (Carlisle, PA: US Army War College Press, 2014).

29 Maura Conway, "We Need a 'Visual Turn' in Violent Online Extremism Research," *VoxPol Blog* (2019), available at https://www.voxpol.eu/we-need-a-visual-turn-in-violent-online-extremism-research/.

30 Jens Seiffert-Brockmann, Trevor Diehl, and Leonhard Dobusch, "Memes as Games: The Evolution of a Digital Discourse Online," *New Media & Society* 20 (2018), no.8: 2862-2879.

31 Scrivens, Freilich, Chermak, and Frank, "Data Collection".

32 Stephane Baele, Katharine Boyd, and Travis Coan, "Lethal Images: Analyzing Extremist Visual Propaganda from ISIS and Beyond," *Journal of Global Security Studies* 5, no.4 (2019): 634-657

33 Stephane Baele, Katharine Boyd, and Travis Coan, "Lethal Images."

34 Charlie Winter, *The Terrorist Image. Decoding the Islamic State's Photo-Propaganda* (New York: Hurst, 2022).

35 Stuart Macdonald and Nuria Lorenzo-Dus, "Visual Jihad: Constructing the "Good Muslim" in Online Jihadist Magazines," *Studies in Conflict & Terrorism* 44, no.5 (2021): 363-386.

36 Sarah Awad, Nicole Doerr, and Anita Nissen, "Far-right Boundary Construction Towards the "Other": Visual Communication of Danish People's Party on Social Media," *British Journal of Sociology* 73, no.5 (2022): 985-1005.

37 Catherine Tebaldi, "Granola Nazis and the Great Reset. Enregistering, Circulating and Regimenting

Nature on the Far Right," *Language, Culture & Society* 5 (2023), no.1: 9-42.

38 Tebaldi, "Granola Nazis".

39 Engstrom, "The Online Visual Group Formation of the Far Right," p.2.

40 Tebaldi, "Granola Nazis".

41 Drew Halfmann and Michael Young, "War Pictures: The Grotesque as a Mobilizing Tactic," *Mobilization: An International Quarterly* 15, no.1 (2010): 1-24

42 Carol Winkler, Kareem El Damanhoury, Aaron Dicker, and Anthony Lemieux, "The Medium is Terrorism: Transformation of the About to Die Trope in Dabiq," *Terrorism & Political Violence* 31, no.2 (2019): 224-243.

43 Lilie Chouliaraki and Angelos Kissas, "The Communication of Horrorism: A Typology of ISIS Online Death Videos," *Critical Studies in Media Communication* 35, no.1 (2018): 24-39.

44 These are often interpreted as "projectilic" images, to use the concept developed by Marwan Kraidy, "The Projectilic Image: Islamic State's Digital Visual Warfare and Global Networked Affect," *Media, Culture & Society* 39, no.8 (2017): 1194-1209.

45 Nicole Doerr, "Bridging Language Barriers, Bonding Against Immigrants: A Visual Case Study of Transnational Network Publics Created by Far-right Activists in Europe," *Discourse & Society* 28, no.1 (2017): 3-23.

46 Nicole Doerr and Beth Gharrity Gardner, "After the Storm. Translating the US Capitol Storming in Germany's Right-wing Digital Media Ecosystem," *Translation in Society* 1, no.1 (2022): 83-104.

47 Nicolò Miotto, "Visual Representation of Martyrdom: Comparing the Symbolism of Jihadi and Far-right Online Martyrologies," *Journal for Deradicalization* 32 (2022): 110-163.

48 Julia DeCook, "Memes and Symbolic Violence: #Proudboys and the Use of Memes for Propaganda and the Construction of Collective Identity," *Learning, Media & Technology* 43 (2018), no.4: 485-504.

49 Roland Bleiker, "Pluralist Methods for Visual Global Politics," *Millennium: Journal of International Studies* 43, no.3 (2015): 872-890.

50 Engstrom, "The Online Visual Group Formation of the Far Right."

51 For instance, Hendry and Lemieux's study of Atomwaffen Division videos only operationalized 12 straightforward categories such as "key symbols", "flag", "infrastructure", or "militant". John Hendry and Anthony Lemieux, "The Visual and Rhetorical Styles of Atomwaffen Division and their Implications," *Dynamics of Asymmetric Conflict* 14, no.2 (2021): 138-159.

52 Jungseock Joo and Zachary Steinert-Threlkeld, "Image as Data," p.12.

53 See Elizabeth Pearson, Joe Whittaker, Till Baaken, Sara Zeiger, Farangiz Atamuradova, and Maura Conway, "Online Extremism and Terrorism Researchers' Security, Safety, and Resilience: Findings from the Field," *VoxPol Reports* (2023).

54 Roderick Hart, "Redeveloping DICTION: Theoretical Considerations," in *Theory, Method and Practice in Computer Content Analysis*, edited by Mark West (New-York: Ablex, 2001): 43–60.

55 Ekaterina Zagarniuk, "What is Computer Vision?" *Medium*, 24 February 2022, https://medium.com/@zagarnyuk.ek/what-is-computer-vision-10c7027e867. This article provides a clear overview of how computer vision works and its main applications.

56 Joo & Steinert-Threlkeld 2022: 17.

57 *ImageNet* accurately classified over a million photos from a challenge corpus into 1000 distinct classes ("leopards", "mushrooms", "pumpkins", "container ships", etc.). See Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Advances in Neural Information Processing Systems* 25, no.2 (2012).

58 See Mario Haim, Marc Jungblut, "Politicians' Self-depiction and their News Portrayal: Evidence from 28 Countries Using Visual Computational Analysis," *Political Communication* (2020), online before print; Amogh Joshi and Cody Buntain, "Examining Similar and Ideologically Correlated Imagery in Online Political Communication," *arXiv* 2110.01183v2 (2022).

59 Yilang Peng, "What Makes Politicians' Instagram Posts Popular? Analysing Social Media Strategies of Candidates and Office Holders with Computer Vision," *International Journal of Press/Politics* 26, no.1 (2021): 143-166.

60 Theisen and colleagues for example offered a method that "ingest images from a social network, apply computer vision-based techniques to extract local features and index new images into a database,

and then organize the memes into related genres" (p.714), and used this method to study over 2 million images circulated on Twitter and Instagram during the 2019 Indonesian presidential elections. See William Theisen, Joel Brogan, Pamela Thomas, Daniel Moreira, Pascal Phoa, Tim Weninger, and Walter Scheirer, "Automatic Discovery of Political Meme Genres with Diverse Appearances," *Proceedings of the International AAAI Conference on Web and Social Media* 15, no.1 (2021): 714-726.

61 Joel Finkelstein, Savvas Zannettou, Barry Bradlyn and Jeremy Blackburn, "A Quantitative Approach to Understanding Online Antisemitism," *arXiv* 1809.01644v1 (2018).

62 Blyth Crawford, Florence Keen, and Guillermo Suarez-Tangil, "Memes, Radicalization, and the Promotion of Violence on Chan Sites," *Proceedings of the Fifteenth International AAAI Conference on Web and Social Media (ICWSM)* (2021): 982-991.

63 Kay O'Halloran, Sabine Tan, Peter Wignell, John Bateman, Duc-Son Pham, Michele Grossman and Andrew Vande Moere, "Interpreting Text and Image relations in Violent Extremist Discourse: A Mixed-Methods Approach for Big Data Analytics," *Terrorism & Political Violence* 31, no.3 (2019): 454-474.

64 These online spaces were identified through both extensive exploration of incel online spaces and snowball sampling and were selected based on both their content and their use of images in posts.

65 Bum Chul Kwon, Ben Eysenbach, Janu Verma, Kenney Ng, Christopher De Filippi, Walter F. Stewart, and Adam Perer Kwon, "Clustervision: Visual Supervision of Unsupervised Clustering," *IEEE Transactions on Visualization & Computer Graphics* 24, no.1 (2018): 142-151.

66 Grimmer, Roberts and Stewart, "Machine Learning for Social Sciences," p.407.

67 Sungwon Park, Sungwon Han, Sundong Kim, Danu Kim, Sungkyu Park, Seunghoon Hong, and Meeyoung Cha, "Improving Unsupervised Image Clustering With Robust Learning," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021): 12278-12287. For a recent tutorial see https://medium.com/@info.martinanto/yolov3-for-weapon-detection-artificial-intelligence-dialling-911-6a4b47ba637c.

68 Grimmer and Stewart, "Text as Data," 280.

69 Grimmer and Stewart, "Text as Data," 281.

70 Joshi and Buntain, "Examining Similar and Ideologically Correlated Imagery," 10.

71 Junyuan Xie, Ross Girshick, Ali Farhadi, "Unsupervised Deep Embedding for Clustering Analysis," *Proceedings of the 33rd International Conference on Machine Learning* (2016).

72 An important point not just because of the aforementioned overfitting issue, but also because of the computational time required to run such models for more epochs.

73 Richard Rogers, "Visual Media Analysis for Instagram and Other Online Platforms," *Big Data & Society* 8, no.1 (2021).

74 The far-right cluster map provided in Appendix 1 is less directly informative because the algorithm had trouble sorting the main type of visual imagery present on the Telegram channels: text-image combinations. Discounting the text-only pictures, far-right Telegram channels appear to heavily picture photos of individuals (Trump, Biden, other US politicians, US military officers, far-right influencers, Nazi leaders, etc.).

75 Similarly, the far-right cluster map in Appendix 1 groups together images of politicians in formal settings (typically wearing suits in front of flags), but gives no indication as to who these individuals are or when members share this type of photographs. The same applies to the cluster grouping faces with texts: it bundles together evidently distinct categories of people (e.g. Nazi Germany officials, US politicians) but is unable to clarify what these categories are and what kinds of text (motivational, defamatory, etc.) accompany them to form multimodal meaning.

76 Zhengxia Zou, Keyan Chen, Zhenwei Shi, Yuhong Guo, and Jieping Ye, "Object Detection in 20 Years: A Survey," *Proceedings of the IEEE* 111, no.3 (2023): 257-276.

77 F. Sultana, A. Sufian, P. Dutta, "A Review of Object Detection Models Based on Convolutional Neural Network," in: *Intelligent Computing: Image Processing Based Applications,* ed. J. Mandal and S. Banerjee, vol. 1157, *Advances in Intelligent Systems and Computing* (Singapore: Springer, 2020). This review and the one cited in the previous endnote offer excellent explanations of how object detection models work, information on the differences between them, and overviews of their steady improvement over the past decade.

78 Zou and colleagues, "Object Detection in 20 Years," p.257.

79 Zou and colleagues, "Object Detection in 20 Years," p.259

80 See for example Sanam Narejo, Bishwajeet Pandey, Doris Esenarro Vargas, Ciro Rodriguez, and Anjum Rizwan, "Weapon Detection Using YOLO V3 for Smart Surveillance System," *Mathematical Problems in Engineering* (2021), doi:10.1155/2021/9975700.

81 Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi Redmon, "You Only Look Once: Unified, Real-Time Object Detection," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016): 779-788.

82 The authors would like to thank the team behind https://www.makesense.ai/, which was used for this task.

The authors would like to extend their thanks to the developers of Ultralytics (https://docs.ultralytics.com/) for their work on these open-source YOLO models.

83 The developers of the COCO dataset, which is the widely adopted benchmark dataset used within the computer vision research and development community, have outlined the issues and extensive work needed to develop such a large data consisting of 2.5 million labelled instances across 328, 000 images; see Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, Piotr Dollar, "Microsoft COCO: Common Objects in Context," *Proceedings of the 13th European Conference on Computer Vision (ECCV)* (2014).

84 The reason behind this low cut-off value was to yield a training dataset that would simply demonstrate how the YOLO model behaves with different numbers of instances for different classes.

85 See https://docs.ultralytics.com/.

86 The pornographic images have been blurred with a white square.

87 For example Engstrom, "The Online Visual Group Formation of the Far Right," or Miotto, "Visual Representation of Martyrdom".

88 Grimmer, Roberts and Stewart, "Machine Learning for Social Sciences," 404.

89 Elanie Rodermond and Frank Weerman, "The Strengths and Struggles of Different Methods of Research on Radicalization, Extremism, and Terrorism," *Studies in Conflict & Terrorism* (2024), online before print: 1.

90 Grimmer and Stewart, "Text as Data," 270.

91 Grimmer, Roberts and Stewart, "Machine Learning for Social Sciences," 403.

92 Bucy and Joo, "Editors' Introduction."

93 Catherine Bouko, *Visual Citizenship. Communicating Political Opinions and Emotions on Social Media* (London: Routledge, 2023).

94 For example Rodermond and Weerman, "Strengths and Struggles," or Bart Schuurman, "Research on Terrorism, 2007–2016: A Review of Data, Methods, and Authorship," *Terrorism & Political Violence* 32, no. 5 (2020): 1011-26.

95 For a review of these assessments, see Schuurman, "Research on Terrorism, 2007–2016."

96 These two observations are documented in Schuurman, "Research on Terrorism, 2007–2016."

97 Baele and Brace, "AI Extremism."

98 Bucy and Joo, "Editors' Introduction."

99 Ryan Scrivens, Joshua Freilich, Steven Chermak, and Richard Frank, "Data Collection in Online Terrorism and Extremism Research: Strengths, Limitations, and Future Directions," Studies in Conflict & Terrorism (2024), online before print: 1-23.

100 Fernandez and Alani, "Artificial Intelligence and Online Extremism," Read also Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman, "Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries," *Frontiers in Big Data* 2, no.13 (2019).

101 All publications offering analysis of incel online content include text but not images, and several have called for this blind spot to be corrected. For example, we read an encouragement to conduct "a study of visual tropes reflecting an endorsement of violent extremism, such as avatar profiles containing pictures of killers or nazi iconography" in Stephane Baele, Lewys Brace, and Debbie Ging, "A Diachronic Cross-Platforms Analysis of Violent Extremist Language in the Incel Online Ecosystem," *Terrorism & Political Violence* 36 (2023), no.3: 382-405.

102 Stephane Baele, Lewys Brace, and Travis Coan, "From 'Incel' to 'Saint': Analyzing the Violent Worl-

dview Behind the 2018 Toronto Attack," *Terrorism & Political Violence*, 33(2019), no.8: 1667-1691.

About

Perspectives on Terrorism

Established in 2007, *Perspectives on Terrorism* (PT) is a quarterly, peer-reviewed, and open-access academic journal. PT is a publication of the International Centre for Counter-Terrorism (ICCT), in partnership with the Institute of Security and Global Affairs (ISGA) at Leiden University, and the Handa Centre for the Study of Terrorism and Political Violence (CSTPV) at the University of St Andrews.

Copyright and Licensing

Contact

E: pt.editor@icct.nl
W: pt.icct.nl

Universiteit Leiden

HANDA CSTPV