Critical simulation as hybrid digital method for exploring the data operations and vernacular cultures of visual social media platforms

Burgess, Jean[a,1], Daniel Angus[a], Nicholas Carah[b], Mark Andrejevic[c], Kiah Hawker[d], Kelly Lewis[e], Abdul Obeid[f], Adam Smith[g], Jane Tan[e], Robbie Fordyce[h], Verity Trott[h], Luzhou Li[h]

a. ARC Centre of Excellence for Automated Decision-Making and Society (ADM+S), Digital Media Research Centre, and School of Communication, Queensland University of Technology
b. ADM+S, School of Communication and Arts, The University of Queensland
c. ADM+S, School of Film, Media and Journalism, Monash University
d. School of Communication and Arts, The University of Queensland
e. Digital Media Research Centre, Queensland University of Technology
f. ADM+S and Digital Media Research Centre, Queensland University of Technology
g. Office of eResearch, Queensland University of Technology
h. School of Film, Media and Journalism, Monash University

[1] Correspondence: jean.burgess@qut.edu.au

Abstract

In this paper we outline and demonstrate the critical simulation approach to understanding the data operations of visual social media platforms. We situate this approach within the field of platform studies and position it as a 'hybrid digital method', before describing its application for descriptive, forensic and speculative purposes in two current research projects: one that uses machine vision combined with mixed-methods qualitative research to explore Instagram's algorithmic visual culture; and one that combines automated data donation and machine vision to explore Facebook's ad targeting practices.

Keywords: Instagram, visual social media, digital humanities, digital methods, algorithmic culture, critical data studies, platform studies.

Introduction

The goal of this paper is to describe the conceptual framework and some applications of the *critical simulation* approach to understanding the data operations of visual social media platforms, including automated content curation, personalisation, and moderation, and focusing particularly on Instagram and Facebook.

Our paper is primarily a contribution to platform studies (Burgess, 2021), a sub-field of digital media, communication and cultural studies that is concerned with the dynamic and co-constitutive relations among platforms' business logics, data operations, and vernacular practices. Recent developments in this field build on the broad corpus of digital methods that has emerged at the intersection of critical, social and technical questions provoked by the internet and social media.

By 'digital methods', we mean more than using digital tools and techniques in the conduct of research. Rather, building and extending on Rogers (2013), we use the term to denote the field of digital media, communication and cultural studies that uses 'the methods of the medium' to study the medium itself, its relations to social and cultural dynamics, and how these relations change over time. As this field (and the digital media environment) has evolved, digital methods have come to integrate computational approaches in a number of ways, and the tools and techniques that enable these approaches have in turn been developed and deployed for descriptive, forensic, and speculative purposes.

The first wave: descriptive digital methods

In earlier phases of web history, digital methods relied on open web data such as hyperlinks to describe patterns in internet use, on which theories and analyses of social processes—like the relationships between environmental organisations implied by links among their websites, for example (Rogers, 2013)—could be built. With the rise of social media, and Twitter in particular, API-based data access saw a flood of studies describing patterns of content, stakeholders, and social networks across a wide range of social science and humanities disciplines. Within digital media, communication and cultural studies, such studies usually tried to integrate the critical study of Twitter's role in mediating the particular forms of communication under investigation: see, for example, Zizi Papacharissi's (2015) work on affective publics, built on the empirical analysis of political hashtags on Twitter; and, relatedly, Bruns and Burgess's (2015) diagnosis of the shift to *ad hoc*, and then algorithmically *calculated* publics, made manifest by the platform's communicative dynamics.

Platformization and the algorithmic turn: forensic digital methods

Since the mid-2000s, however, the platformization (Helmond 2015) of the web (which, among other things, involves the marketization, enclosure and attempted control of data by platform companies) has made access to the data required for systematic, descriptive analysis more difficult. Even more recently, the related turn to algorithmic culture (which occludes the data operations of platforms) have increasingly necessitated new hybrid digital methods (Burgess 2021; Burgess and Bruns 2020) as researchers attempt to understand the relationships between platform interfaces, operations, and cultures of use; or, to adopt a popular metaphor, to get a look inside the 'black box'.

The 'black box' metaphor has long been used to describe systems that operate in or act on the world in consequential ways, but whose inner workings and decision-making processes remain obscured from view. The term has been widely adopted in media and cultural studies to highlight problems with the opacity surrounding platforms' data operations, particularly with respect to search, recommendation, and content moderation, where the processes of automated choice- and decision-making are both unclear and consequential for society (see for example Pasquale, 2015). Where these concerns are operationalised as methods (as in the algorithmic audit, discussed below), they have increasingly taken what we might call a *forensic* turn. That is, forensic digital methods involve the adoption of systematic investigative techniques to generate evidence and develop theories about what platforms are doing with user data, and how they are making decisions about what content to show, hide, or remove, and to whom, often by inferring data operations from the relationship between inputs and outputs.

The most prominent model for forensic digital methods approaches to the study of platform data operations is the algorithmic audit (Sandvig et al., 2014), usually motivated by hypotheses and critical frameworks around algorithmic bias and/or questions of fairness with respect to automated platform governance and content moderation. For example, Noble (2018) mobilises an input/output approach to study the racist implications of algorithmically driven autosuggestions and search engine results; Witt et al. (2019) bring a legal feminist perspective to investigate Instagram's black box processes of content moderation through a study on user-generated images that depict women's bodies, finding that up to 22 percent of images removed for violations of Instagram's content policies are potentially false positives (Witt, et al., 2019, p. 587).

Towards speculative digital methods

Building on this work in the digital methods tradition, in this paper we argue that critical platform and data studies needs to develop new digital methods approaches which move beyond static forensic approaches such as algorithmic audits. Rather than attempting to divine or reverse engineer the workings of the 'black box' of platform data operations, critical simulation involves a form of dynamic 'grey box' testing that invites creative, explorative investigation, using advanced computational techniques, guided by critical and qualitative questions. In doing so, critical simulation projects develop software tools and interfaces that can be used far more widely for descriptive analysis of large corpora of platform artefacts, and as an element of mixed-methods qualitative research into these cultures.

Paper outline

In the remainder of this paper, we home in on the specific research challenges for critical platform studies presented by visual social media platforms, including their entanglement with advertising, their relationships to the broader AI ambitions of platforms; and, of course, ongoing challenges around data access. We outline the critical simulation approach and its contribution to hybrid digital methods/critical data studies, with descriptive, forensic, and speculative applications.

We then outline two examples of computational methods integrated with critical conceptual frameworks and in support of qualitative research, incorporating both close textual analysis and participatory methods with vernacular experts on the cultures under study. First, the Image Machine and its applications for studying everyday promotional cultures of Instagram; and second, the browser plugin and associated research protocols for studying 'dark' advertising on Facebook.

We show how these computational tools can be integrated with qualitative methods to empower the descriptive, forensic and speculative study of algorithmic digital media cultures, and conclude by outlining the next steps in our research.

Specific challenges of studying visual social media platforms

The vernacular production of images of everyday life is integral to the commercial model of platforms like Instagram (Burgess, 2015, Banet-Weiser, 2012, Abidin, 2016) and its parent company Facebook. In addition to contributing signals to the data pipelines that feed collaborative filtering and targeted advertising, participants do the everyday work of incorporating commercial experiences, objects and logos into the stories they tell about themselves. A critical, and often under-examined, aspect of this relationship is that the everyday production, not only of behavioural signals like 'likes' and comments, but also of these images, is critical to training the machine vision classification of the platform itself (Carah and Angus, 2018; Andrejevic, 2019).

Trade press reporting and patents can help us to understand more about how machine vision is linked to the commercial interests, and particularly the advertising model, of platforms. Unlike development blogs, patents do not track actual implementations made by platforms, rather, they are discursive and performative: they signal what platforms say they are *intending* to do, in turn influencing the market. Whilst this is a speculative view, in conjunction with development blogs and trade-press it offers an insightful look into the public positioning if not the underlying motives of major platforms with respect to machine vision.

In 2018, Facebook's engineering and AI department, which develops systems for both Facebook and Instagram, labelled image recognition as an integral pillar to Facebook's growth and success (Paluri et al. 2018). To illustrate their focus on these systems the platform consistently highlights the large scales that their models can train and evaluate. Facebook stated in 2018 that their systems can now scan upwards of 50 million images, with the goal of increasing this into the billions (Paluri et al., 2018). Facebook also showcases moves to improve the ways these systems function, in a 2020 post the platform details how 'transformer technology' is being incorporated to advance their computer vision efficiency (Touvron et al., 2020). Furthermore, a self-supervised computer vision system from Facebook named SEER (SElf-supERvised), claims to be able to learn from any image without the need for any human input or curation (Goyal et al., 2021). These examples showcase the push from Facebook to produce extremely efficient and scalable machine vision processes, which can operate with little to no human intervention.

Three patents by Facebook, Inc. from the past decade illustrate a clear push to create machine vision systems directly to help advertise branded content to users (Systrom & Cole, 2020; Systrom & Cole, 2020; Mitchell et al., 2011). This objective is most explicitly shown within a 2020 patent labelled 'Method for Selectively Advertising Items in an Image', which describes methods to employ 'Image-classification algorithms' to recognize 'products, locations and brands in images and automatically auction them to businesses and brands as an advertising opportunity' (Systrom & Cole, 2020). The goal is for Facebook to utilise machine vision technology to automatically identify certain 'key' objects in user posts and then advertise related products to these users. The trade press surrounding Facebook also reiterates this desired application. In response, Razzaq (2019) raises concerns around user labour and the way these developments might transform the advertising industry, potentially leading to those images containing branded content being prioritised in terms of news feed curation over those that do not (Razzaq, 2019).

Snapchat also applies machine vision technology to better advertise products to users. The camera company that owns Snapchat, Snap Inc. lodged a patent in 2018, describing a form of image recognition technology to better identify 'media items' or branded content within user images, and creating 'engagement scores' which can be analysed (Davis et al., 2018). This helps Snapchat evaluate which products or items are most popularly shown within their app. This data is extremely valuable to advertisers and the platform, who can use it to determine the effectiveness of certain products/brands, with certain users, and in specific locations (CBInsights, 2017). Additionally, a 2017 patent named 'Emotion recognition in video conferencing', notes the platform's ambition to link the facial expressions of a user with previously detected objects within their posts, to target users with certain products at the right time (Shaburov & Monastyrshyn, 2017). For example, if a user looks 'tired', and has previously posted an image with coffee present, they would be more likely to be served an advertisement for coffee-related products.

Smaller start-up companies also work with major media platforms to facilitate machine vision systems. The media intelligence company GumGum was arguably one of the first companies to introduce machine vision advertising strategies in 2008 (Arrington, 2008). Through machine vision and deep learning tech, GumGum scans the layouts and content of platform news feeds or home pages to evaluate the effectiveness of advertising placements. A 2015 patent by the company describes methods for systems 'to associate advertisement data with content within social platforms' (Mir & Tanz, 2015). This patent demonstrates just how long-standing is the objective to apply machine vision systems within social media environments for the purpose of advertising.

The machine vision systems described by platform companies in their patents and trade press both exploit and run counter to everyday cultural practices. For platforms, understanding images' meanings and representational qualities is not as important as learning how they can be used to nudge behaviour and shape cultural tastes. In the 'fantasies' of platforms and marketers as represented by patent and business literature, practices of cultural expression like creating and sharing images are processed into data that reveal patterns; and these patterns enable platforms to generate 'highly personalized worlds that are algorithmically emptied of irrelevant choices' (Darmody and Zwick, 2020). 'Targeting' is 'used less by larger advertisers, who tend to rely on their own lists of users or algorithmic replication of past campaigns (Cotter et al., 2021). Instead of targeting based on predefined criteria, advertisers increasingly use the algorithmic models of platforms to 'optimise' audiences. In these models, images are not only part of practices of narration and meaning-making, but also feed open-ended pattern recognition and prediction processes. However, within critical data studies and digital media research, the tools and methods for the critical oversight of such systems, let alone the critical assessment of their relationship to the vernacular cultures of platforms, are as yet in their infancy. Critical simulation is our contribution to this effort.

Positioning critical simulation as hybrid digital method

A critical move in studying the algorithmic cultures of visual social media platforms is to move beyond the interface, to explore deeper layers of the platform's computational stack, and contextualise the data operations at each layer in relation to platforms' business models and cultures of use. In the vernaculars of software engineering, stacks are assemblages of technologies that support specific flows and processing of data throughout various input, processing, and output layers, with specific interdependencies between layers of the stack, and requirements for their application and use. From a critical platform studies perspective, especially in the context of platforms involving social interaction and user-contributed content, we understand these flows of data as being inextricable from the social relations and cultural practices of social media users. Undertaken as part of mixed-methods approaches to investigating the data operations of platforms in their sociocultural contexts, critical simulation involves the development of a technology stack that is analogous to some facet of the platform under

investigation, and can typically include a data layer, algorithmic layer, and various visualisations and statistical outputs from these layers to provide greater insights into the internal operations and affordances of the system, and in turn to enable critical, qualitative analysis.

In the first instance, critical simulation involves building open-access software that uses algorithms comparable to those used by the platform of interest. The idea here is not to directly emulate the platform, but rather create an information processing pipeline that replicates many of the same technologies and computational moves as the platform can be assumed to use. In doing so we can begin to expose how user data could be processed in a transparent manner, and to build hypotheses around the various factors that may impact the classification of user materials.

Meanwhile, software engineers have their own approaches to investigating algorithmic operations, albeit framed in the language of 'testing' – specifically, white box, black box, and grey box testing of the relationship between a piece of software's inputs and its outputs. In the case of white box testing, full access to the underlying computer source code is provided to the tester. In black box testing no access is provided to the source code, only the program's input and output functions, making the tester's level of software access closer to what an end user might expect. Grey box testing is, as the name suggests, a hybrid of white and black box testing. In grey box testing the tester may not have access to source code, but will have more of an idea about the specific algorithms and data structures used by the program.

In the context of critical social science and humanities research that seeks to understand the algorithmic operations of digital media platforms, in some cases, we do have details of at least the class of algorithm, data sources and structures, and other information, positioning some aspects of platforms' data operations as grey-box systems. And it is in this context where critical simulation can play a role in assisting in the demystification and public understanding of these systems.

Critical simulation, then, deploys exploratory computer science approaches within the context of critical inquiries about platform operations, develops practical tools for data analysis, and in turn generates provisional findings that can be explored within mixed-methods qualitative research. Thus, alongside the app walkthrough, platform biography and others, we can frame critical simulation as a hybrid digital method (Burgess, 2021).

Introducing the Image Machine

Our *Image Machine* is designed to support the critical simulation of Instagram's machine vision systems. It comprises an open-access machine vision toolbox that implements analogous algorithms to those in use by Instagram (informed through trade press and grey literature), to allow researchers to experiment with, interrogate and analyse the ways that images associated with a particular aspect of Instagram's vernacular cultures might be classified and/or clustered by the platform itself, and the data logics that might be at work in these operations – in this sense, it supports *speculative* digital methods grounded in an understanding of industry practices. The system also integrates visualisation techniques that can be used by humanities researchers to critically investigate the outputs of different deep learning models that we have reason to believe are used by the major visual social media platforms, and so in this sense it supports *forensic* approaches to the critical study of machine vision. And finally, its visualisation (e.g. of image clusters) enable qualitative research (using both close textual analysis and participatory methods) to be guided by broad explorations of the content of large datasets, in this context operating supportively as a *descriptive* digital method.
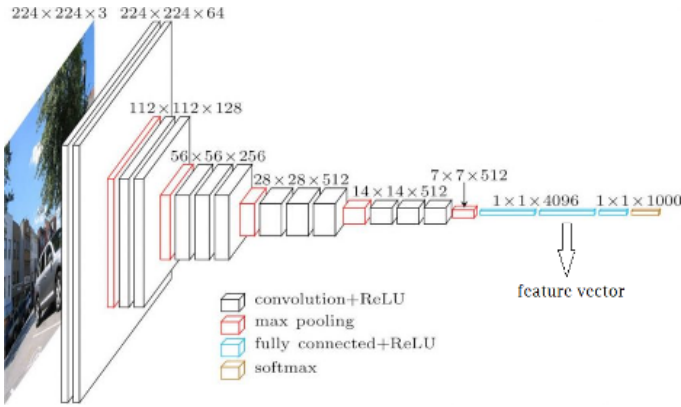
Machine vision algorithms

Machine vision algorithms can be roughly categorised into two general classes: 'supervised' and 'unsupervised'. Supervised approaches involve the training (read: adjustment) of algorithm parameters to create a mapping between data inputs (images) and pre-defined categories/features as outputs. In image classification this might involve labelling images as to the presence or absence of visual objects and to then train a computational model to classify previously unseen images into these known classes as accurately as possible.

Supervised approaches are deductive, predominantly involving a corpus of input training data – including corpora such as ImageNet - being fed through the classifier and the output judged for accuracy against predetermined 'labelled' features of the input data. Depending on the approach, adjustments are made in specific ways to tune the classification algorithm to improve these predictions, often through an iterative process repeated over multiple training cycles until a stopping criterion is reached, such as a slowdown in improvement in classification accuracy. As part of this training process, supervised algorithms are often checked against hold-out test data to determine whether the trained classifier can generalise beyond the initial training dataset, or if alternatively it has overfit the training data, meaning that it may perform poorly on other unseen examples taken from the intended domain of application. Once trained, these supervised algorithms are somewhat compact and transportable, meaning that they can be shared and reused in many different capacities often as part of a software library or web service, without any further need for, or dependency on the initial training data. The ease of reuse of pre-trained models brings with it significant risk of misapplication of these models beyond the circumstances and contexts for which they were originally intended and trained.

In contrast, unsupervised classification approaches involve the use of algorithms to inductively group input data samples based on latent data features/properties, often where this input data is characterised by a high number of dimensions. Unlike supervised classification there is no formal train/test adjustment cycle, as these algorithms aren't trying to 'learn' pre-assigned categorical mappings. Rather unsupervised algorithms seek to cluster/group/divide data, in the case of music this may be according to latent frequency characteristics, in text it could be along latent word statistics such as the relative frequency of specific word collocations. In some approaches media objects may belong to many different clusters, as in the case of hierarchical clustering, or only one cluster, as in the case of k-means clustering. The distinguishing characteristic of unsupervised classification is that resultant clusters aren't tuned or adjusted to try to map a predetermined reality (i.e. does this image contain a bicycle? Is this a reggae song?), instead resulting patterns and classifications are interpreted post-hoc based on the presence of shared or contrastive features.

Data processing: embedding and clustering

To interrogate and experiment with both supervised and unsupervised machine vision approaches, the data processing pipeline of the Image Machine consists of an *embedding* module, a *clustering* module, and a *visualisation* module. Each module is customisable, and in some cases there are dependencies between elements in each module i.e. specific visualisations for specific clustering algorithms, or clustering algorithms that require a specific form of embedding. Image datasets are ingested first into the embedding module, which is where we implement and can make adjustments to existing machine vision algorithms. The embedding module is responsible for mapping input images into numeric vectors or other outputs. The clustering layer uses the vectors and other outputs from the embedding module to perform clustering operations on these 'embedded' images. Clustering helps us understand how images are 'seen' by the various algorithms in the embedding layer by revealing associative connections between the

Figure 1. The VGG16 convolutional neural network. Our unsupervised approach takes the third to last output layer, to perform an Image2Vec embedding process on input images.

images. The visualisation layer allows the various image cluster and embedding information to be visualised and interrogated.

One machine vision algorithm we have implemented in the embedding module is the popular VGG16 algorithm (Simonyan & Zisserman, 2014). VGG16 has been used by Facebook as part of their 'detectron' framework (Detectron, 2018), and is a supervised image classification model that employs a convolutional neural network architecture. VGG16 comes pre-trained using images from the ImageNet database to identify the presence of 1000 discrete object categories. While VGG16 is a supervised algorithm, meaning the final layer maps to a set of discrete object categories, our initial use of it here utilises the second-last layer of the multi-layer VGG16 network. This is a common practice in embedding, which allows the algorithm to be used to map input images from their original resolution and colour depth (RGB) into a uniform 4096-element numeric representation, rather than

sending the image through to the final layer of discrete 'supervised' categories (see Figure 1).

Given the focus on embedding, we are not immediately interested in the labelled outputs - as we would if the VGG16 network were being used in standalone setting (to perhaps determine the presence of specific discrete pre-trained objects in an image) - rather we are interested in interrogating how this algorithm maps images into similar areas of a 4096-dimension feature space. The use of VGG16 in this way is similar to that of an autoencoder, which is an unsupervised neural network often used to perform dimensional reduction or other data re-mappings. In the case here, the VGG16 network maps an image that is made up of 224 x 224 x 3 = 150,528 data elements, to a feature vector of 4096 elements. This 4096 element vector can then be used as the basis for clustering operations in the subsequent clustering module.

The clustering module is considered as an unsupervised stage of the Image Machine because we don't try to pre-assign images to clusters, instead we allow the clustering algorithms to discover latent clusters of images based on the embedding vectors. Similar to the embedding module there are many possible clustering approaches which can provide different lenses through which we can interrogate the associative connections between images that are made by the embedding module. Some clustering algorithms may need to compare all images to all other images (full pairwise) such that images have membership to multiple clusters that are organised hierarchically, while others like k-means clustering create centroids (points in the hyperspace) that are iteratively adjusted through comparisons with the full image set such that images only belong to one distinct cluster.

Data visualisation

There are many standard approaches to visualising clustered data, and the approach we have initially adopted for the hierarchical cluster is a

Figure 2. The visualisation module allows clustered image data to be visualised and interrogated. In this example the VGG16 embedding is combined with a hierarchical clustering, operating over a corpus of images from Instagram.
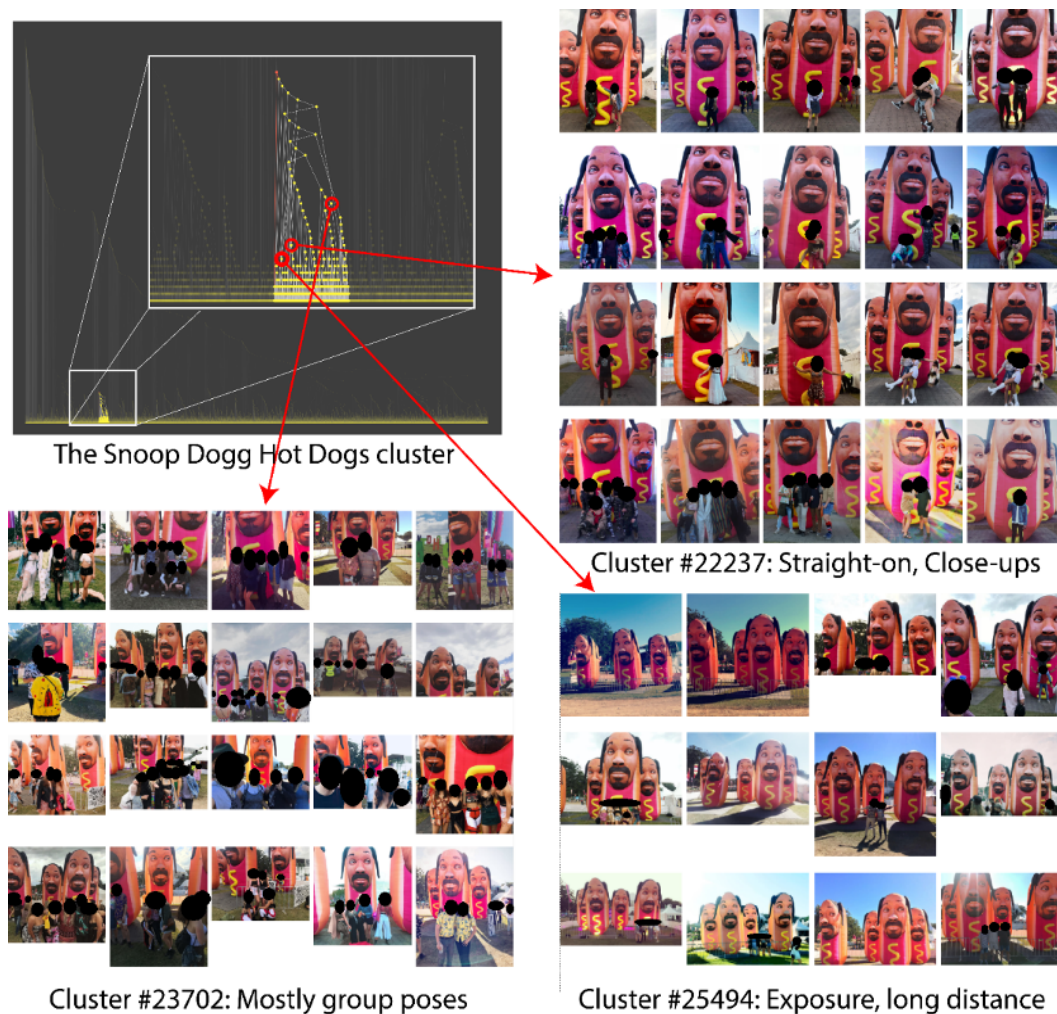
dendrogram. In a dendrogram, each node represents a single cluster, with edges connecting from the immediate descendent and antecedent clusters in the hierarchy. The software implementation utilises the D3.js visualisation framework, so it is straightforward to extend this approach in order to deploy alternative visualisations, such as force-directed layouts, or other tree structures. The dendrogram layout reads from top to bottom, with the upper left most node representing the entire image collection, proceeding down and rightward towards the bottom which contains each individual image cluster. The visualisation is incorporated into a user interface that allows an analyst to interrogate any single cluster, and view sample images from this cluster, or the entire contents (see Figure 2). Basic statistics such as the median cluster distance, a measure of the homogeneity of images in cluster, are also reported.

The clustering model does not prescribe category/cluster names, leaving such names and descriptions up to researchers to label (if desired) via iterative data-led interpretations. In the example provided in Figure 2, the highlighted cluster (#25448) contains photographs of food. Exploration of adjacent nodes might reveal that this branch of the dendrogram is primarily comprised of similar food-focussed photography.

In our recent study of Instagram material gathered from the Australian music festival 'Splendour in the Grass' (Carah et al., forthcoming), we found that art installations, commercial branding, specific site locations, bands, and other visual ephemera were clustered, despite these containing visual objects and features that the underlying supervised VGG16 network had not been explicitly trained on. As one example, images of the *Snoop Dogg Hot Dogs* art installation was clearly not something that had been used in training the VGG16 network, nor were the hip hop artist (Snoop Dogg) nor foodstuff (American style Hot Dogs) that inspired this creative work. Despite not being trained explicitly on these objects, the machine vision approach was able to associate all images of this object, taken from a variety of poses, compositions and lighting conditions, given that they are located at a similar location of the dendrogram (see Figure 3).

Using the Image Machine for descriptive, forensic, and speculative analysis

We are now deploying the Image Machine in a multi-year study of how platforms' machine vision techniques might be operationalised within the advertising model and promotional culture of platforms like Instagram (Carah & Angus, 2018), drawing on a range of everyday

Figure 3. Adjacent clusters of Instagram images all contain the same distinctive art installation, but are grouped according to similar photographic techniques and compositions.

visual subcultures (ranging from lifestyle sports like skateboarding, creative practices like fashion cultures, graffiti and classroom design, and everyday practices like home decorating and home cooking), which involves integrating critical simulation with an ethnographic approach to those cultures.

We intend to use critical simulation and associated ethnographic approaches to explore each of these cases in descriptive, forensic and speculative ways. To take one example: in the case of everyday material practices like home cooking and home decoration we begin by assembling a collection of images of these practices posted using public hashtags. For example, generic hashtags like #homecooking, #kmartmumsaustralia, #thehomeedit or hashtags related to particular cookbooks, celebrity chefs or food, homewares and kitchen brands.

A *descriptive* approach employs the machine vision system to identify patterns of expression and composition in the Instagram images associated with a particular topic. For example, the machine vision clustering might identify the patterned dominance of particular meals, kitchen items, or cookbooks; an informed, context-aware cultural analysis could explore these patterns and how they change over time. Or, it might identify dominant modes of composition like the framing of kitchens and dining tables or the plating of meals, providing an insight into the aesthetics of 'instagrammable' domestic spaces. In this way, the Image Machine clustering operations and resulting visualisations might provide an excellent starting point – something to 'think with' – for a manual thematic or cultural analysis of the Instagram (sub)culture in question, simply by ordering and presenting the data to the researcher, enabling them to build a 'thick' description from the Image Machine's relatively 'thin' ones.

A *forensic* approach to these images could examine how different algorithms cluster and classify the images. We might employ different algorithms to undertake unsupervised clustering. For instance, running the model once using VGG16 and then again using VGG19. This might reveal that one model is more likely to organise clusters based on the everyday objects that featured in its training data, or that one model is more weighted toward the presence of people or facial expressions. This would enable us to forensically investigate how different models, their training data, histories, weightings and assumptions shape the organisation of images. We could also undertake a supervised classification of images to forensically investigate how accurate machine vision models are at labelling features in images. This would enable us to examine the capacity of a machine vision model to identify brand logos or distinctive objects like cookbooks or branded kitchen items. A forensic approach could also associate the clusters created by the machine vision model with metadata about engagement with the images, such as the number of likes or views. Or, to sequence the emergence of practices over time. This could enable us to model how image features might correlate with, and be assumed to predict, user engagement. This would enable us to investigate how particular practices of expression by influencers and celebrity chefs might shape the everyday forms of expression by ordinary users.

A *speculative* approach uses the machine vision model as the basis for critically imagining the possible applications of its capacity to process images by platforms and advertisers. This requires us to situate the outputs of the model in relation to the everyday and promotional culture, and the advertising and recommendation model of the platform. We would conceptualise and envisage how platforms could use the capacity to cluster images based on their features in identifying patterns, and feeding those patterns into the operations of their advertising and recommendation engines. We would also explore how users develop an algorithmic imaginary, a set of vernacular theories, about how machine vision operates around them as they express themselves and represent their everyday practices on platforms. Future speculative work could use images of these everyday practices as

training data for creating generative adversarial networks that would explore the capacity of platforms and advertisers to create wholly synthetic images (such as customised, simulated advertisements) on the basis of our everyday expressions.

In combination with each of these descriptive, forensic and speculative approaches we will undertake interviews with home cooks and decorators, or influencers and creators, or social media managers to discuss their practices of cooking and decorating using Instagram, to undertake exploratory exercises where we give them copies of images posted to Instagram and ask them to describe and organise them, and then to show them visualisations created through the machine vision model. Our purpose here is to involve participants in these everyday cultures as co-interpreters of their practices and the legibility of those practices to machine vision systems. We would explore with participants both how the machine vision systems describe their cultural practices (including highlighting the profound limitations of machine classification); and how, despite its flaws, machine vision might be used by platforms to intervene in, and shape, those cultural practices.
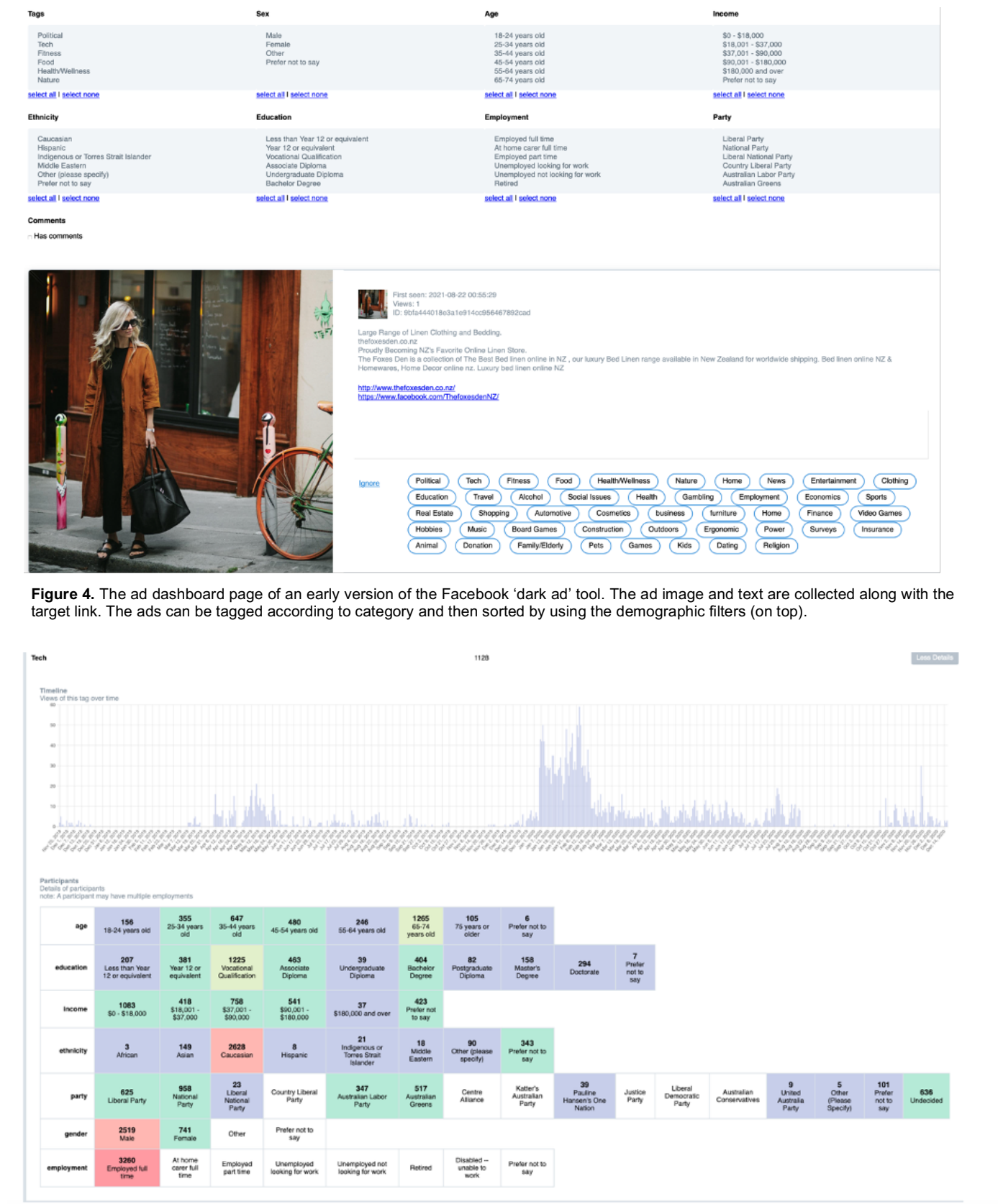
The Facebook 'dark ads' project

A second project combines data donation from volunteers with an ad clustering analysis generated by the Image Machine to gain insight into online ad targeting. Initially piloted at Monash University, inspired by work undertaken at NYU in collaboration with ProPublica, the project is now hosted by the ARC Centre of Excellence for Automated Decision-Making and Society and known publicly as the Australian Ad Observatory - see https://www.admscentre.org.au/adobservatory/ for details.

As in other forms of forensic analysis, the project employs an output-focused approach as a means of addressing the fact that we do not have access to Facebook's targeting algorithms – or to the targeting information stipulated by advertisers. Since we are unable to explore the intermediary mechanisms, or even the intentions of the platform or those who use it for promotional purposes, we focus on the different types of advertising received by different demographic groups. We know this to be an important area of social concern because of the recent history of Facebook paying a legal settlement after making it possible to target ads for housing, jobs, and credit by race, age, and gender – all protected categories for this type of advertising (Yurieff, 2019).

Such activities fit into a longer history of predatory advertising, discrimination, and stereotyping in advertising (see, for example, David, 2020; Tuckson, 1989). Whereas these practices had some level of public visibility in the mass media era, the shift to online advertising marks a tectonic shift in the ability to hold advertisers accountable. Online ads are ephemeral, copious, and, in most cases, only visible to those to whom they are targeted. In the case of online job ads, it is difficult for individuals to determine whether they are being singled out on the basis of race, gender, or age, because they do not know who else has seen the ad – and if they happen to be excluded, they may be unaware that the ad even exists.

The Facebook 'dark ads' project seeks to address these challenges to accountability by inviting volunteers to install a browser extension that collects all the ads they encounter in their news feeds while using Facebook on a desktop or laptop computer. Volunteers are invited to provide basic demographic information when they install the tool, allowing for an analysis of how ads are targeted according to any combination of these categories (see Figure 4).

Figure 4. The ad dashboard page of an early version of the Facebook 'dark ad' tool. The ad image and text are collected along with the target link. The ads can be tagged according to category and then sorted by using the demographic filters (on top).

Figure 5: A sample breakdown of the demographic distribution of technology ads in our pilot study. The frequency table above indicates when the ads were collected, reflecting the fact that the study was conducted in two phases and some people did not disable or remove the extension.

The prototype version of the dashboard enabled manual tagging of ad categories, providing visibility into how particular types of ads are targeted to different demographic groups, and continuing work is exploring ways to semi-automate this tagging process. A particular ad category, such as 'tech ads,' once selected, generates a profile demonstrating how that category was distributed across participants in the study. This function can be adjusted to quantify unique views or unique users (to correct for different amounts of time spent online). In our pilot study of 160 participants, for example, people who self-identified as male were more than three times more likely to receive a technology ad than those who self-identified as female, regardless of the gender settings of their actual Facebook profiles; we did not have any participants who self-identified as non-binary (see Figure 5).

The project demonstrates one approach to providing visibility into online ad targeting, but it is limited by the number of volunteers who are willing to install the tool and by the ability to process the data captured. With our updated version of the plugin in the Australian Ad Observatory we are proactively recruiting a larger and more diverse cohort of volunteers, with a specific focus also on Indigenous Australian Facebook users. Clearly, we cannot approach the level of information available to Facebook, not least because we collect only limited data, voluntarily provided, about users, but also because of the different scale at which we are operating.
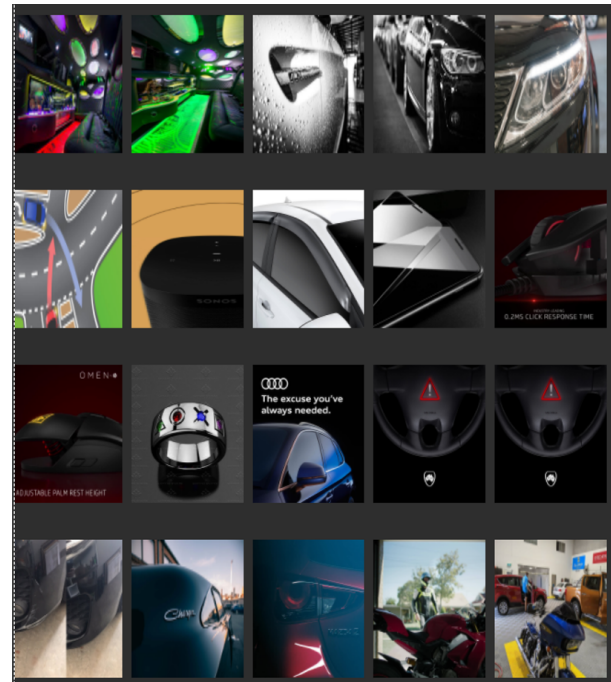
However, the tool does have the benefit of making ad targeting outcomes visible and of capturing ads that might not make it into the Facebook ad library. It also allows us to ask questions that relate to public concerns about Facebook but that Facebook may not be interested in asking or answering publicly, such as how patterns of ad targeting might reinforce harmful stereotypes or amount to predatory or discriminatory forms of commercial messaging.

Our system does not directly emulate Facebook's ad targeting process from end to end, but it does *simulate* some of Facebook's data operations with respect to ad targeting, and thereby provides insights into the patterns that emerge from and feed back into this process. We know, for example, that Facebook's automated systems target ads based on the performance of other similar ads – and the patterns we see are the results, in part, of this feedback.

The other challenge posed by the 'dark ads' project is how to analyse the data we collect at scale. Within the pilot work, even a small number of participants (100) for a delimited time (one month) generated tens of thousands of ads. Implementing the project at a larger scale will generate more data than can be reasonably tagged manually or analysed without the assistance of automated systems. We are working toward developing automated forms of content analysis including Optical Character Recognition and logo detection using the YOLO Object detection framework (Redmon et. al., 2016). We have also already drawn upon the Image Machine to automatically cluster and sort the results from our pilot project.
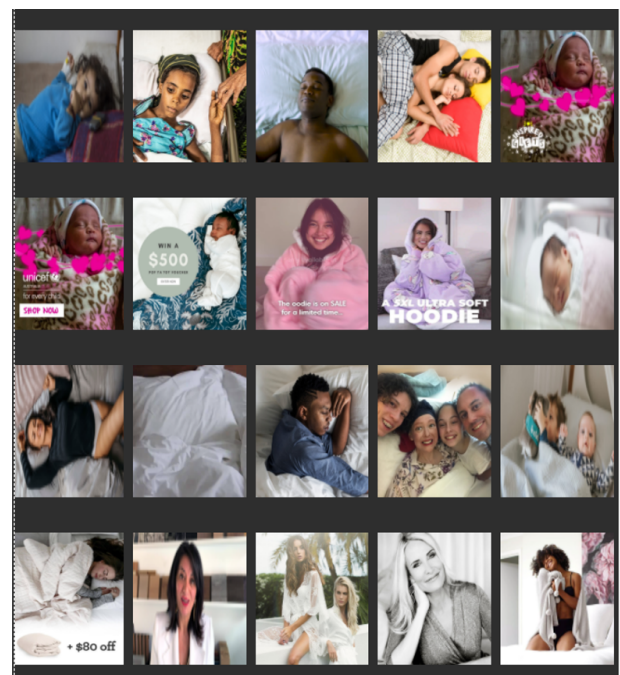
The first step in the analysis was to create a set of clusters from all of the images collected by the browser extension. In our pilot study, the classifier identified 66 clusters in our data set. Although the size of these varied from just a few images to scores of them, we could see from each cluster, thanks to the data visualization, which audience demographic characteristics were associated with it. We note that this is a matter of interpretation, but that this is an important aspect of assessing how associations and stereotypes might be reproduced by targeted advertising. For example, the sleek-looking car ads cluster illustrated in Figure 6 has a decidedly masculine skew whereas the 'Sleeping Related' cluster in Figure 7 had a very strong female skew.

The advantage of the Image Machine for this type of unsupervised cluster analysis is that it allows us to gain insight into how images

Figure 6. The automatically generated "Sleek Cars" cluster had a strong skew toward users who self-identified as male

might reinforce forms of cultural stereotyping. It also simulates the way in which Facebook itself compares new ads to previous ones in order to optimise at targeting. We know, for example, that it compares not only the text of new ads to previous ones in order to optimise performance, but also the images (Ali, et al., 2019). We can then use the tool to get a sense of which ads might be deemed similar to one another based on their images. With further enhancement we could also enlist the tool for automated forms of category tagging by using it to recognise brand logos and icons, and thus to group together ads for

Figure 7. The automatically generated "Sleep related" cluster had a strong skew toward users who self-identified as female.

the same products. All of these aspects of the 'dark ad' tool can then be used to inform discussion with volunteer participants who are able to access the dashboard in order to view their own ad targeting history. They are able to use an anonymized user number to see all the ads collected from their Facebook news feed during their participation in the project. This allows us to enlist the 'scrollback' method (Robards & Lincoln, 2017) to get their own impressions of how they are being targeted – as part of the qualitative analysis component of the research project.

Conclusion

Critical simulation methods such as those enabled by our Image Machine and Facebook Dark Ads plugin can help inform critical understanding of how machine learning algorithms and automated decision-making systems mediate, curate, shape and intervene in the ongoing automation of social and cultural worlds (Andrejevic, 2020, p.159). But a crucial piece of this puzzle, in a context where platform cultures are co-created by users (Burgess & Baym, 2020) is the question of how, in turn, users contribute to (whether willingly or not), creatively negotiate, and push back against machinic operations.

As discussed above in our walkthrough of the Instagram project, our next move is to examine how participants understand and creatively navigate cultural settings that are partly organised by machines. In the case of Instagram, bringing the inner workings and outputs of machine vision algorithmic decision making into dialogue with Instagram users can allow us to better examine and reflect on how machines operate on everyday visual cultures (Carah & Angus, 2018). Critical simulation allows us to ask what kinds of new relationships are forming between vernacular practices and machine classification, including how everyday participation in visual social media also helps to train machine vision systems, and how human and nonhuman processes are interrelated in the production of digital culture – as well as finding spaces of hope among the everyday data practices and vernacular knowledges that transcend or escape the protocols of machine vision and algorithmic curation (Burgess, Albury, McCosker & Wilken., 2022). Critical simulation, as a hybrid computational and qualitative research method, allows us to ask what kinds of new relationships are forming between vernacular practices and machine classification – including the ways that machine operations are in tension or completely at odds with vernacular practices and knowledges.

In the case of Facebook advertising, the 'dark ads' project allows us to move beyond speculating about the ways in which targeting might enable discrimination, predatory advertising, and stereotyping to test the competing claims in this area, and especially to explore concrete examples. Previous approaches using similar tools by researchers at NYU and investigative journalists at ProPublica and the Guardian have focused on identifying political advertising that might otherwise have flown under the radar, but have not collected demographic information about participants in order to identify empirical outcomes of ad targeting. Such information, while not comprehensive, could bolster initiatives for greater transparency as well as forms of regulation that envision alternative systems for shaping our news and information ecosystem.

From a critical perspective, the approaches described in this paper provide a response to the dramatic informational and computational asymmetry of platform media, which appear to structure public and private interaction and participation in ways that remain largely opaque to users, while at the same time potentially failing to deliver on their own hype with regard to omniscient knowledge and perfectly targeted advertising.

It is important to continue to develop and explore new methods that can aid in observing and understanding the operations of platforms, and to use these observations to reflect on the commercial logics and structures that shape and govern the news and media environment more broadly. There is a direct link between an examination of the everyday uses of these structures and our shared ability, as a society, to consider how we would like them to be regulated and, if necessary, transformed.

End Matter

Author Contributions and Notes

Burgess coordinated the paper, led the writing and editing, and contributed key conceptual elements relating to hybrid digital methods, including the 'descriptive, forensic, and speculative' schema; **Angus** contributed to the writing and key conceptual elements relating to algorithmic systems, machine vision, and critical simulation, and led the development of the technical methods described in the paper; **Carah** contributed to the writing and key conceptual elements relating to critical simulation, platform and advertisers uses of machine vision and descriptive, forensic and speculative approaches to studying Instagram; **Andrejevic** contributed to the writing, focusing on the description, theory, and history in the Facebook 'dark ads' section; **Hawker** contributed to the writing and made conceptual contributions to understanding how advertisers and platforms use machine vision; **Lewis** contributed writing and conceptual elements relating to forensic digital methods, and provided research assistance including the collection of Instagram data; **Obeid** led the technical development of the Australian Ad Observatory; **Smith** developed the technical infrastructure used to collect Instagram data, and the initial prototype and software architecture of the Image Machine.; **Tan** further developed the Image Machine visualisation tool as an extension to the prototype built by Smith, and has extended the Instagram data collection methods; **Fordyce**, **Trott, and Li** all assisted in conceptualising the Facebook dark ads project and in analysing the data gathered by the ad collector.

The authors declare no conflicts of interest.

References

Abidin, C. (2016). Visibility labour: Engaging with Influencers' fashion brands and #OOTD advertorial campaigns on Instagram. *Media International Australia* 161(1): 86-100.

Ali, M., Sapiezynski, P., Bogen, M., Korolova, A., Mislove, A., & Rieke, A. (2019). Discrimination through optimization: How Facebook's Ad delivery can lead to biased outcomes. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW): 1-30.

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media and Society* 20(3): 973–989.

Andrejevic, M. (2013). *Infoglut: How too much information is changing the way we think and know*. Routledge: London.

Andrejevic, M. (2020). *Automated Media*. New York: Routledge.

Arrington, M. (2008). GumGum Launches New Image Licensing Platform. *TechCrunch*. Retrieved from: https://techcrunch.com/2008/02/13/gumgum-launches-new-image-licensing-platform/

Banet-Weiser, S. (2012). *AuthenticTM: The politics of ambivalence in a brand culture*. New York: NYU Press.

Benjamin, R (2019). *Race After Technology: Abolitionist Tools for the New Jim Code*. Cambridge: Polity Press.

Bivens, R. (2015). Under the hood: The software in your feminist approach. *Feminist Media Studies* 15(4): 714–717. https://doi.org/10.1080/14680777.2015.1053717

Bivens, R. (2017). The gender binary will not be deprogrammed: Ten years of coding gender on Facebook. *New Media and Society* 19(6): 880–898. https://doi.org/10.1177/1461444815621527

Bivens, R., & Haimson, O. L. (2016). Baking Gender Into Social Media Design: How Platforms Shape Categories for Users and Advertisers. Social Media and Society, 2(4). https://doi.org/10.1177/2056305116672486

Bratton, B. H. (2016). *The Stack: On Software and Sovereignty*. MIT press.

Bruns, A. & Burgess, J. (2015). Twitter hashtags from ad hoc to calculated publics. In Rambukkana, N (Ed.) *Hashtag publics: The power and politics of discursive networks*, pp. 13-27. New York: Peter Lang.

Burgess, J. (2015). From 'Broadcast yourself' to 'Follow your interests': Making over social media. *International Journal of Cultural Studies*, 18(3): 281-285.

Burgess, J. (2021). Platform studies. In Cunningham, S. & Craig, D., eds. *Creator Culture*. New York: NYU Press.

Burgess, J. (2020). *Twitter – A Biography*. New York: NYU Press.

Burgess, J. & Bruns, A. (2020) Digital Methods in Africa and Beyond: A View from Down Under. *African Journalism Studies*, *41*(4), pp. 16-21. https://eprints.qut.edu.au/207865/

Burgess, J., Albury, K., McCosker, A. & Wilken, R. (2022, forthcoming) *Everyday Data Cultures.* Cambridge: Polity Press.

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data and Society* 3(1): 1–12. https://doi.org/10.1177/2053951715622512

Carah, N., & Angus, D. (2018). Algorithmic brand culture: participatory labour, machine learning and branding on social media. *Media, Culture & Society,* 40(2), 178–194.

CB Insights. (2017, May 10). What Snap's patents say about their efforts In augmented reality, wearables, facial recognition, and more. *CB Insights*. Retrieved from: https://www.cbinsights.com/research/snapchat-patents/

Cotter, K., Medeiros, M., Pak, C., & Thorson, K. (2021). "Reach the right people": The politics of "interests" in Facebook's classification system for ad targeting. *Big Data & Society*, *8*(1), 2053951721996046.

Crawford, K. (2018). 'Trouble with Bias' The 31st Annual Conference on Neural Information Processing Systems (NIPS 2017) Keynote Presentation. Retrieved from https://www.youtube.com/watch?v=ggzWIipKraM

Darmody, A., & Zwick, D. (2020). Manipulate to empower: Hyper-relevance and the contradictions of marketing in the age of surveillance capitalism. *Big Data & Society*, 7(1).

Davis, J.F. (2020). Representation Matters: An Illustrated History of Race and Ethnicity in Advertising. *Advertising & Society Quarterly 21*(3), doi:10.1353/asr.2021.0002.

Davis, B., Sellis, P., Azmoodeh, F., Ahmed, S. & Cohen, D. (2018). *System to track engagement of media items.* (Patent No. US10165402B1). Retrieved from: https://patents.google.com/patent/US10165402B1/en?q=computer+vision&assignee=Snapchat&oq=Snapchat+computer+vision

Detectron. (2018, January 23). Retrieved from https://github.com/facebookresearch/Detectron

Diakopoulos, N. (2014). Algorithmic-Accountability: The investigation of Black Boxes. Report, Tow Center for Digital Journalism, Columbia University, USA.

Diakopoulos, N. (2015). Algorithmic accountability: Journalistic investigation of computational power structures. *Digital Journalism* 3(3): 398–415.

Diakopoulos, N. (2016). Accountability in algorithmic decision making. *Communications of the ACM* 59(2): 56–62.

Diakopoulos, N. & Koliska, M. (2017). Algorithmic transparency in the news media. *Digital Journalism* 5(7): 809–828.

Eubanks, V. (2018). *Automating inequality: how high-tech tools profile, police, and punish the poor* (First Edition.). St. Martin's Press.

Farocki, H. (2004). Phantom Images. *Public,* 29.

Gillespie, T. (2014). The Relevance of Algorithms. In: Gillespie T, Boczkowski P and Foot K (eds) *Media Technologies: Essays on Communication, Materiality, and Society*. Cambridge, MA: The MIT Press (pp. 167-193).

Gillespie, T. (2017). Algorithmically recognizable: Santorum's Google problem, and Google's Santorum problem. *Information, Communication & Society* 20(1): 63–80.

Goyal, P., Caggiano, V., Joulin, A. & Bojanowski, P. (2021). SEER: The start of a more powerful, flexible, and accessible era for computer vision. *Facebook AI.* Retrieved from: https://ai.facebook.com/blog/seer-the-start-of-a-more-powerful-flexible-and-accessible-era-for-computer-vision/

Helmond, A. (2015). The platformization of the web: Making web data platform ready. *Social Media + Society* 1(2): https://doi.org/10.1177/2056305115603080

Highfield, T., & Leaver, T. (2016). Instagrammatics and digital methods: studying visual social media, from selfies and GIFs to memes and emoji. *Communication Research and Practice*, 2(1): 47-62.

Mir, A. & Tanz, O. (2015). *Systems and methods for electronically managing content*. (Patent No. US9135637B1). Retrieved from: https://patents.google.com/patent/US9135637B1/en?assignee=GumGum&oq=GumGum

Mitchell, J., Odio, S, & Garcia, D. H. (2011). *Computer-vision content detection for sponsored stories*. (Patent No. US9135631B2). Retrieved from: https://patents.google.com/patent/US9135631B2/en?q=(Facebook+image+recognition+advertising)&assignee=Facebook&page=1

Moe, H. (2019). Comparing Platform "Ranking Cultures" Across Languages: The Case of Islam on YouTube in Scandinavia. *Social Media + Society.* https://doi.org/10.1177/2056305118817038

Noble, S.U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: New York University Press.

Packer, J. (2013). Epistemology not ideology or why we need new Germans. *Communication and Critical/Cultural Studies*, 10(2-3), 295-300.

Paluri, M., Mahajan, D., Girshick, R. & Ramanathan, V. (2018). Advancing state-of-the-art image recognition with deep learning on hashtags. *Facebook Engineering.* Retrieved from: https://engineering.fb.com/2018/05/02/ml-applications/advancing-state-of-the-art-image-recognition-with-deep-learning-on-hashtags/

Pasquale, F. (2015). *The Black Box Society: The Secret Algorithms That Control Money and Information*. Cambridge: Harvard University Press.

Pilipets, E., & Paasonen, S. (2020). Nipples, memes, and algorithmic failure: NSFW critique of Tumblr censorship. *New Media & Society*. https://doi.org/10.1177/1461444820979280

Polack, P. (2020). Beyond algorithmic reformism: Forward engineering the designs of algorithmic systems. Big Data and Society, 7(1). https://doi.org/10.1177/2053951720913064

Razzaq, A. (2019). Facebook's New Image Recognition Algorithm Can Scan Your Picture For Advertising Opportunities. *Business2Community*. Retrieved from: https://www.business2community.com/facebook/facebooks-new-image-recognition-algorithm-can-scan-your-picture-for-advertising-opportunities-02202720

Burgess et al., 2021.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 779-788).

Robards, B. & Lincoln, S. (2017). Uncovering longitudinal life narratives: scrolling back on Facebook. *Qualitative Research* 17(6): 715-730.

Seaver, N. (2019). Knowing Algorithms. DigitalSTS, 412–422. https://doi.org/10.2307/j.ctvc77mp9.30

Shaburov, V. & Monastyrshyn, Y. (2017). Emotion recognition in video conferencing. (Patent No. US9576190B2). Retrieved from: https://patents.google.com/patent/US9576190B2/

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Systrom, K & Cole, A. (2020). *Method for Selectively Advertising Items in an Image*. (Patent No. US20140279039A1). Retrieved from: https://patents.google.com/patent/US20140279039A1/en

Touvron, H., Douze, M., Massa, F., Sablayrolles, A. & Jegou, H. (2020). Data-efficient image Transformers: A promising new technique for image classification. *Facebook AI*. Retrieved from: https://ai.facebook.com/blog/data-efficient-image-transformers-a-promising-new-technique-for-image-classification/

Tuckson, R. V. (1989). Race, sex, economics, and tobacco advertising. *Journal of the National Medical Association*, *81*(11), 1119.

Turow, J. (2012). *The Daily You: How the New Advertising Industry is Defining Your Identity and Your Worth*. Yale University Press: New Haven.

Van den Oord, A., Dieleman, S., & Schrauwen, B. (2013). Deep content-based music recommendation. In proceedings of Neural Information Processing Systems 2013 (NIPS 2013), 2643-2651.

Witt, A., Suzor, N., & Huggins, A. (2019). The rule of law on instagram: An evaluation of the moderation of images depicting women's bodies. *University of New South Wales Law Journal* 42(2): 557-596.