Designing a Moral Compass for the Future of Computer Vision using Speculative Analysis

Michael Skirpan
University of Colorado
Boulder, CO
michael.skirpan@colorado.edu

Tom Yeh
University of Colorado
Boulder, CO
tom.yeh@colorado.edu

Abstract

In this paper we discuss and analyze possible futures for technologies in the field of computer vision (CV). Using a method we have coined speculative analysis we take a broad look at research trends in the field to categorize risks, analyze which ones are most threatening and likely, and ultimately summarize conclusions for how the field may attempt to stem future harms caused by CV technologies. We develop narrative case studies to provoke dialogue and deeply explore possible risk scenarios we found to be most probable and severe. We arrive at the position that there are serious potentials for CV to cause discriminatory harm and exacerbate cybersecurity issues.

1. Introduction

Computer vision (CV) techniques are at the epicenter of excitement and progress related to recent developments in deep learning; specifically, convolutional neural networks [36, 29, 30, 64]. Concomitant with a surge in the success of machine learning systems is unprecedented access to new datasets [14, 6, 19, 23]. There is no end in sight for this growth in promise and applications. The massive availability of image data coming from commercial sources such as Flickr, Instagram, and Facebook, and the dispersed use of ubiquitous and smart camera systems has locked us into a future where our living image will constantly be monitored, captured, processed, and used to generate inferences.

Without a doubt, combining advanced machine learning with troves of image data is likely to aid human causes such as health monitoring [37] and accelerate efficiency in areas such as archiving [17] and traffic analysis [66]. However, with the blinding light of promise glistening, we must be careful not to miss that there are consequences and dangers to allowing these applications to run amok.

Over the past few years, we have seen many red flags waved that should caution researchers to how deep learning and CV may go wrong. Machine learning techniques have been critiqued for their ability to inherit bias and create discriminatory results on tasks that may have chilling consequences [34]. MIT researcher Joy Buolamwini began the Algorithmic Justice League after discovering that a common face recognition software failed to work on black faces [15]. Tech giant, Google, was forced to dial back image captions as their software regularly identified photos of black people as Gorillas [4]. Further, the threat of unwarranted or unfair surveillance is greater than ever as police forces are deploying facial recognition algorithms on massive scales with further threats to discrimination and injustice [2]. And these concerns are just the tip of the iceberg as IoT cameras have proven to be easily exploited [53] and computer vision techniques have developed that undermine privacy [45] and security [63].

Seeing these promises and concerns growing hand-in-hand, we must adopt techniques for comprehending and communicating these risks and steering technology toward a future society we all want. In short: we must figure out how to guide rapidly developing fields, such as computer vision, with a moral compass.

In this paper, we raise ethical questions and ultimately speculate on possible futures being offered by certain CV technologies that may have unfair and dangerous consequences for our society. To structure our discussion we use a method we are developing called *speculative analysis*. Our approach brings together various areas of research such as the study of risk perception [55], speculative design/fiction [13, 7], and future studies [40, 57] in order to garner foresight, generate representations of alternative futures, and analyze competing ethical factors relevant to technology and policy decision-making. The goal of the paper is to provoke ethical discussion among CV researchers and practitioners and argue for the position that risks around security and discrimination caused by CV technologies have a high likelihood of negatively impacting our future society.

We start by offering a categorization of risks based on a critical reflection of recent papers emerging from CV research, specifically using past years' work in CVPR and the arXiv. The categories are drawn from an interpretive method [40] of scanning current events in order to relate the ethical concerns of the real world to publication trends in the literature. These categories are used to provide structure for a further speculation into a series of hypothetical future scenarios.

Each scenario is meant to ground and characterize risks from our categories. We consider different combinations of vulnerable populations, technology implementations, and harms in order to obtain risk characterizations akin to how a traditional risk analysis would begin [21]. Given the amount of uncertainty around the future applications of CV technology, we stray from traditional risk analysis, as has been suggested by other risk researchers [31], in order to promote thinking from a broader societal and ethical vantage. We drill into each speculative scenario by analyzing the likelihood of it occurring and comparing its risk factors. The conclusions of our *speculative analysis* are used to justify the position that security and discrimination risks are the most novel and threatening and demand serious attention by the CV community over the coming years. In relation to our position, we offer up two short pieces of original speculative fiction meant as communicative and educational tools to provoke conversation between members of the CV community on these issues.

We conclude by discussing lessons one might derive from these speculative fictions, what further work could be done to improve our analysis, and how the field of CV may move forward utilizing such considerations.

2. Categorizing Risk Factors in Computer Vision

2.1. Bounding Our Considerations

Prior to characterizing and categorizing risk factors, a few framing assumptions and definitions must be made. Given that our discussion is highly normative and interpretive, it is crucial to state the lens we are applying throughout. Much like a risk assessment, we are not simply providing a summary of science, but attempting to enhance practical understanding to guide the future decision-making of a particular group [21]. While it is possible to consider CV ethics from a business, organizational, research, or policy lens, each would demand different framings and evaluations. For the consideration of this paper, our target audience is the community of CV researchers and practitioners. That is, we will analyze risk from a lens relevant to those deciding how to design, develop, and form best practices for new computer vision technologies. Thus in lieu of other possible perspectives, we will focus on how systems emerging from the community of researchers and engineers may allow for certain futures and thus risks felt by the broader society. We choose this framing due to our agreement with Langdon Winner that ”artifacts have politics” [59]. That is, while many of the ethical situations and risks we will discuss have other organizational and application dimensions, it is our view that the production of new CV systems themselves and the possibilities they provide structure the ethics of all downstream considerations. We take the position that if researchers and engineers are not careful to build norms about what they create, the market and the law will not adapt fast enough to prevent risks from being borne onto the public.

This brings us to further delimit our ethical questions; specifically, what we are calling a risk. As risk researchers have established, the first part of defining a risk is to decide which consequences are included [26]. For the sake of this paper, we will define these consequences as harms or hazards that computer vision technologies may inflict on the public. Again, this is distinct from other definitions of risk, such as empirical risk embedded in certain methodological choices [47]. The assumption here is that the CV community cares to assess the societal impacts it may create and review practical considerations on how to avoid dangers and public detriment. Thus, we will limit ourselves to risks that enact a specifiable harm to an individual or group who might interact with a CV technology.

2.2. Categorizing Trends

For the purpose of structuring our discussion, we elaborate on five categories of risk that have CV-specific correlates: *privacy violations, discrimination, security breaches, spoofing and adversarial inputs, and psychological harms*. These categories were derived from a subjective and critical reflection of research papers and ignored other general technology risks such as job loss and error/edge cases. Here we go into detail defining each category and tracing out the correlates from the CV literature and current events that justify its relevance to the conversation.

Privacy Violations: This risk category is meant to cover all ways in which CV applications may lead to a third-party gaining unintentional or undisclosed private information about user. This may include, unwarranted surveillance, inferring information that was undisclosed, or de-anonymizing images. The potential for privacy issues appears pervasive given work currently emerging in CV research such as inferring health metrics from social media images [35] or de-anonymizing blurred images [45]. Further, due to continued progress in facial recognition abilities [10, 50, 33, 28, 9] the presence of any passive camera or image found online could easily lead to identification and potentially a privacy-violating inference. Technology ethics researchers Kate Crawford and Jason Schultz have termed the class of privacy violations that come from unanticipated inference as predictive privacy harms [22]. As our ability to make inferences from videos and images expands, so do the possibilities of diminishing trust and causing predictive privacy harms by inferring unintended and, potentially, consequential private information such as health informatics, uniquely identifying someone who did not choose to post a photo or video, or pinpointing a person's location.

Discrimination: Codified by our laws in place via Title VII of the Civil Rights Act, Title IX of the Higher Education Act, and the ADA, discrimination harms occur when someone receives unfair treatment due to their identity such as race, gender, class, or sexual orientation. Much like humans, the capacity to discriminate is further alive in machines. An undeniable trend in CV is the heightened use of machine learning models, specifically convolutional neural networks [36]. The promise of machine learning is paralleled by the difficulty in making sure the resulting models are fair. Broadly speaking, the fear of bias and discrimination within AI and machine learning has become a topic of the day. Within CV, MIT researcher Joy Buolamwini has found that biased training samples have led to facial recognition models that do not work on black or other minority faces [16]. Further, research on age, race, and gender image classification continues to progress [39]. There is even work attempting to replicate models that can identify female attractiveness from a male viewer [61]. With racial and gender opportunity gaps being a continued problem of our time, technologists must not ignore how they may objectify or exacerbate these issues. Especially concerning are reports that racial bias is already showing up in mass facial recognition software used by police officers [32, 25].

Security Breaches: A large umbrella of risks imposed by CV technologies are varieties of security vulnerabilities. That are ways in which the presence or misuse of cameras or CV systems allow access to guarded information or systems. We classify a broad spectrum of attacks under this category. Propelled by CV innovations, the ubiquity of camera systems has opened up vulnerabilities any time proper security precautions have not been taken. Recently, massive attacks against CCTV cameras in Washington DC allowed up to 70% of security cameras in the region to be compromised [58]. Separately, ubiquitous cameras took a part in a large-scale IoT attack against DNS servers as botnets compromised hundreds of thousands of devices to be used in a one-off DDoS [53]. Further, researchers have shown that cameras can be used to steal information, such as passwords, off of filmed screens [63]. Without appropriate security, mobile cameras, home monitoring systems, web cams, and even calibration systems for critical systems have the potential to be co-opted by adversaries.

Spoofing and Adversarial Input: Broadly defined, this category, in the scope of CV, are adversarial attacks that attempt to get automated systems to react confidently to inputs while generating incorrect results. CV systems that are used for fraud detection, liveness detection, act as a security barrier, or have social consequences may be threatened by an adversary who understands the system and can game it. There has already been a thread of research showing the ability to exploit deep neural nets [42, 56], soliciting high-confidence predictions for humanly-unrecognizable images. Other research has proven that these adversarial inputs are not simply laboratory scenarios, but expose a reality of real-world vulnerabilities [38, 27]. Researchers have already proven that this problem extends beyond well-understood systems. By targeting common CV tasks, such as segmentation and detection, it is possible to create systems that can generate adversarial examples against arbitrary blackbox CV systems [62].

Psychological Harms: Unlike other harms resulting from a CV technology's function, psychological harms are related to the wider effects created by ubiquitous cameras and passive monitoring. Having a world where personal devices, CCTV, drones, satellite images, and social media imagery are omnipresent and potentially smart (ie, actively making inferences) gives the impression of unending surveillance. This may lead to a constant state of stress and anxiety, and perhaps lead people to make social choices, such as not attending a protest, based solely on the fear of scrutiny or exposure. There is already a history of research showing that workplace monitoring leads to employees feeling more stressed [1]. Further work has shown that surveillance is likely to diminish people criticizing the government [18] or their ability and willingness to escape oppression [51].

3. Speculative Analysis

Relating trends in CV to categories of harm avails broad areas for ethical discussion; however, it does not give insights into which trends may be most worrying and where we should pay special attention. In order to get at these practical conclusions, we first must unpack these broad trends into smaller components for analysis. Our unit of analysis is the scenario, which consists of a specific technological arrangement causing a harm to a population. The goal here is to entertain a wide range of scenarios to tease out the ones most worthy of deeper consideration. Utilizing scenarios for rapid, low-cost evaluation of technology has a deep history within design [7], future studies [48], and HCI research [44]. They are commonly used to test boundaries around norms [44], engage users in design processes [20], and enable analysis of future conditions that would otherwise be encountered with high uncertainty [57].

Our motivation was to speculate on scenarios that would offer concrete representations of our abstract risk categories and generate a wide space for comparing different contexts and consequences. As other researchers have suggested [40], understanding the future requires an integrated cultural and technological analysis as well as a readiness to consider negative aspects often glossed over when reporting new research. Given that this is a speculative task of abductive reasoning, we attempted to limit our final analysis to what we considered our "best guesses." That is, we began by generating a massive list of roughly 50 scenarios which we first cut down to about 30 by identifying scenarios that did not have obvious technological feasibility given the current state of research. Then we sent our shortened list to 4 practicing computer scientists, asking them to identify the most far-fetched ones. Filtering down from this feedback, we ended with the list in Table 1.

Throughout Table 1, you will find several scenarios of how discrimination may leak into trained models, particularly CNNs. One scenario we discuss is that a training dataset containing an undiscovered bias gets passed around and used for varying commercial applications to only later discover bias. Given the difficulty in interpreting deep neural nets, if found too late, it could be impossible to fully remove the bias from the trained model. An escalated scenario relates the trend of police using facial recognition to identify suspected criminals. As is being discussed in current events [54, 2], facial samples used by police disproportionately sample African Americans due to historical bias in policing and crime. This could lead to a predictive policing system that uses threat scoring or, even further out, perhaps an autonomous security drone that monitors the public for criminals and has deep biases to suspect African Americans are criminals. Ignoring this possibility may objectify and exacerbate the very problem of prejudice we seek to eradicate in our society.

Separately, we consider scenarios containing security breaches in IoT cameras and surveillance systems. While this problem crosses CV and cybersecurity, we focus on it due to the elevated attractiveness of incorporating CV techniques into larger integrated systems. If a large enough botnet was successful, we could see internet outages that could harm online and business infrastructure for large amounts of time. An even bigger concern would be if distributed IoT networks became a method to pass along more harmful viruses, much like the Stuxnet worm [65], searching for access points into vulnerable infrastructure such as the power grid or broadband systems. These possibilities may warrant a reconsideration of best practices regarding how online cameras are deployed and how they are integrated into larger technical systems.

Postulating ways spoofing attacks could turn awry, we consider scenarios where driverless cars could be attacked to trigger highway collisions by placing a carefully selected object in the visual range of the driverless vehicle. Separately, we could imagine CV being used to discover environmental hazards such as oil spills. If a malicious company wanted to cover up the event, they may tamper with the visual features of the hazard, say spraying a color-changing chemical onto the surface of the spill, to avoid detection by a known CV system.

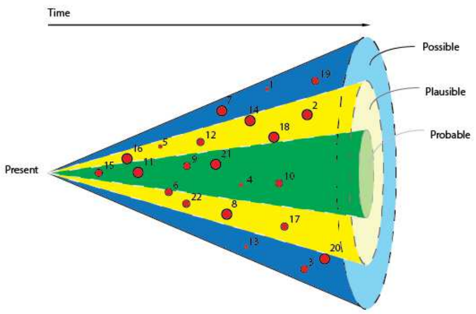

Figure 1. Likelihood categories of possible futures. The size of the dot relates to how big of a population would be affected by the harm and the distance from the present relates to how far away we believe the scenario to be from the present.

Another scenario includes a privacy issue that may come with CV indiscriminately processing mass online photography. One way this could go wrong is if photos are posted online without someone's consent that then get processed, tagged, and finally associated with a profile that costs them a job or harms their personal life. Similarly, we could imagine CV used by insurance companies to better assess the health of its applicants using available image data. This may allow a photo that was never considered relevant to health care to cause a spike in someone's insurance premium or even show evidence of a pre-existing condition that was unknown or untreated.

3.1. Likelihood Analysis

Though each scenario may relate to some possible future, if framed as a Bayesian question $P(Scenario|Technology)$, one would not give each equal likelihood of occurring. As a proxy for a Bayesian question, we took each scenario and categorized it into either possible (meaning it's technically feasible, but easily avoidable), plausible (meaning it's feasible, hard to avoid, but would take a very malicious actor), or probable (meaning it's feasible, hard to avoid, and already on the path to occurring). Using an adapted visual aid referenced in speculative design [7], we show how each of these scenarios ranked in Figure 1. We believe discrimination, large-scale security breaches, and damage to democracy through psychological harm to be among the most probable concerns. Discrimination risks ranked probable due to the popularity of CNNs along with emerging evidence of bias in the data that will likely be used for future training [32, 25] and evidence that discrimination is already occurring in other machine learning systems [34]. Security breaches and psychological concerns were also ranked as probable due to their relevance to current events in

#	Scenario	Technology	Population	Harm
1	Health insurance premium is increased due to inferences from online photos	Biometric inference	Any individual	Privacy
2	People will not attend protests they agree with due to fear of recognition by cameras and subsequent punishment	Face Recognition	Public	Psychological
3	Person is unfairly denied entry into public location due to visual scan at door	Image classification	Minority community	Discrimination
4	Person denied job for photo they did not post online that was de-anonymized by CV.	Face recognition	Any individual	Privacy
5	Security guard sells footage of public official typing in a password to allow for a classified information leak	Key stroke or screen inference	Any individual	Security Breach
6	Job candidate is denied job because classifier of attractiveness was used within a model.	Social psychology classification from image	Women	Discrimination
7	Environmental hazards tracked by aerial imagery are missed because company tampers with visual appearance of a pollutant	Machine learning	Public	Spoof/Fraud
8	Automated public transportation that uses visual verification systems is attacked causing a crash	Driverless cars	Public	Spoof/Fraud
9	Xenophobia leads police forces to track and target foreigners	Facial recognition and image classifiers	Minority populations	Discrimination
10	Police unjustly search and arrest people of color due to criminality inferences.	Human inference from images	Minority populations	Discrimination
11	Online infrastructure is brought down through IoT attack using hacked cameras	IoT cameras	Public	Security Breach
12	Programmer uses third-party CV model to create eye tracking tool that, at release, does not work for Asian faces	Eye tracking	Minority population	Discrimination
13	Autonomous security system using CV, incorrectly detects object as weapon, leading to unjust attack or arrest	Object detection	Anyone	Error
14	Corporate ethics spiral as whistleblowers are deterred by workplace monitoring	Camera surveillance	Public	Discrimination
15	Automatic captioning leads to thousands of offensive captions on public photos	Image captioning	Any individual	Discrimination
16	Anonymized health dataset used for CNN training gets de-anonymized by adversary, revealing health info of millions	CNN and de-anonymization	Public	Privacy
17	Recreational drones for extreme sports video popularizes and incidentally captures videos of children in public spaces	Drone cameras	Children	Privacy
18	Death of private life - people must assume all matters of life may be used against them in work, court, etc	Ubiquitous camera systems	Public	Psychological
19	Disaster response dictated by aerial imagery ends up sending all first responders to rich neighborhoods because classifier uses inferred value of property	Image classifiers	Low SES Populations	Discrimination
20	Automated weapon is triggered to attack innocent people due to adversarial attack against visual processing system	Thread modeling from images	Public	Spoof/Fraud
21	Mass IoT network is used to pass a virus along and deliver into public infrastructure system, taking down portion of power grid by creating timed surges	IoT Cameras	Public	Security Breach
22	A popular image dataset gets used to train dozens of commercial CNN applications, and is discovered to have a major bias in it that disadvantages minority groups	CNN	Minority Group	Discrimination

Table 1. Twenty-Two Risky Scenarios Used for Analysis

cybersecurity [49] and pervasive conversations around the loss of privacy in our society [46].

3.2. Plotting Risk Factors

The next part of our process involved plotting these scenarios along dimensions of *uncertainty* and *severity*. We modeled these factors from a risk perception study [55] published in 2005 to rank risks as perceived by different groups of the public. Since our focus is on future harms to the public from CV technologies, we saw these dimensions as most relevant to codifying our perceptions of the 22 scenarios.

To create a metric for *uncertainty* we considered the following factors: observability (is the harmful effect easily observable?); newness (is this a new risk or one society has long faced?); known exposure (does a person know they were exposed to the risk?); scientific knowledge (is the risk well understood by scientists?). The less known and observable, the newer and more difficult to infer exposure, the more positive the value for uncertainty. The metric of *severity* was structured by a separate set of factors: controllable (do practitioners have a lot of control over the risk?); detrimental (are the harms common or detrimental to the population involved?); scale (do the harms occur at a large, global or small, individual scale?); risk to future generations (are the effects lasting burdens on future generations or quickly addressable?); mitigation (is the risk easily mitigated or difficult?). The less controllable, more detrimental, larger scale, more of a future burden, and harder to mitigate all contributes to a more positive value.

We gave values to this 2D metric based on facts about harms we know through the news, mitigation tactics published by field experts, and how much of a damage the final harm wages (ie., embarrassment < financial loss < physical harm < detriment to societal functions). While we believe our position is a good starting provocation, thorough follow-up studies could be done allowing expert and non-expert populations to weigh in on how they rank and compare risks along these same metrics. Using the above questions and assessment criterion, we constructed the plot appearing in Figure 2.

We will elaborate on how we arrived at some of our highest ranking risks (ie, top right quadrant of Figure 2). Scenario 2 - people afraid of protesting because of surveillance and face recognition - was treated as a vast concern. Not only is it very difficult to observe the actual psychological distress that could cause this, it may occur over a longer

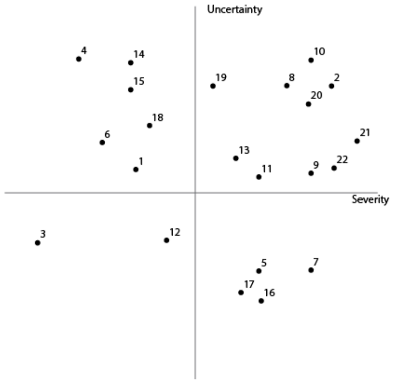

Figure 2. Uncertainty vs. Severity for 22 CV Risk Scenarios

period of time as various punishments accumulate. Also, it could vary drastically depending on who is in power and how laws progress around information sharing. As Frank Pasquale discusses in The Black Box Society, Occupy Wall Street already suffered from some of these conditions as Wall St. banks, unknown to activists, gave security camera footage to the FBI, allowing certain protesters to be identified and targeted [46], likely with the aid of face recognition software. Further, once this change occurred it would impact the ability of future generations to change the status quo making corrupt regimes even more powerful.

Scenario 10 - police unjustly searching and arresting people of color due to a bias in visual analytics done on surveillance and camera monitoring systems - was ranked as both severe and uncertain. Uncertain because a single instance of a person targeted as a threat by a visual monitoring system would not necessarily augur this deeper issue. It may take years of injustice and expert assessment of the systems to fully comprehend the risk. In the meantime, the consequences to human lives would be severe and it could breed further distrust between communities, destabilizing social foundations.

Scenario 21 - IoT cameras carrying along a virus hoping to pass it to vulnerable infrastructure - was assessed as a severe concern. Given the difficulty in controlling security vulnerabilities multiplied by the instances of devices that continue to be networked online creates an exorbitant concern. On the other hand, we did not rank the uncertainty so high. While the attack vectors may grow in size and it's hard to tell if a system is infected, the idea that networked infrastructure needs strict regulation is not new. The use of networked devices to monitor and control infrastructure has increased over decades allowing cybersecurity experts plenty of time to consider attacks. A well-planned attack in this realm could involve multiple CV-based attacks, first passing a virus, then exploiting a CV verification system, or even spoofing a biometric monitor for a security officer.

4. Narrative Case Studies

Narrative and fiction are commonly used constructs for research and exploratory purposes within HCI [13, 11, 52] and design theory [7]. The added value of narratives over simply adumbrating scenarios or analyzing fail modes of systems is that they create a context that is better able to represent social or political conflicts [12]. That is, locating the consequences and frictions of engineering decisions is often difficult in purely technical descriptions of systems where accuracy and efficiency, system dynamics, and usability are often analytic constructs that lend to objective solutions. However, when we consider harm to people, we need richer depictions that allow us to consider thorny matters such as social norms, notions of justice and fairness, and trust. As DiSalvo argues, fictional examples offer opportunities for interrogation and challenge [24]. A good technological narrative should structure a place for discussion, disagreement, and ethical deliberation among experts. While we do not have the space to construct narratives for every scenario nor elaborate on details of future worlds, we offer up two provocative flash fictions (Figures 3 and 4) to deconstruct, analyze, and add to the repertoire of conversation among CV researchers

In the Scenario 10 narrative (Figure 3), we see two men discussing an event where one was arrested due to a police confrontation instigated by a camera system that identified him as a threat. The implication is that society has committed itself further to the utility of smart camera systems, trusting the inferences they make to guide police response and create efficiencies in physical security. Presumably connected to a vast database of faces and operating with an AI model that evaluates likelihoods of someone's intentions, the inferences made by visual data have allowed discrimination to move into a more objective realm. As stated by the protagonist, he was targeted for reasons that were discovered, much later, to be related to a racial bias within the system. We implicitly understand that he must have been an innocent person meaning it is unlikely the system was connected to any human-in-the-loop overseer who may have been able to redirect the police who responded. Further, he informally suggests that the history of racial bias in policing should have made programmers anticipate severe biases in any system trained by that data.

How far off is this kind of scenario? And how dangerous should we see it to our society? The idea of predictive policing is already on the rise. Data-driven approaches and machine learning applications are currently being tooled to predict crime [41, 32, 3]. Guessing the likelihood of recidivism [8] and setting bond [34] are further becoming areas where computer science is at work, and already signs of racial bias are showing up [25]. With computer vision research emerging that attempts to predict criminality of a face [60], and the many advances of object detection and

Ray exited the correctional facility a free man. The long walk between the brick-and-mortar structure and the gated entry was lined by swiveling cameras. Sentries were replaced by intelligent observation systems - Watchful Eyes, as the policing community called them. Being the same system that got Ray into trouble six months ago, he and his brother couldn't shake a looming sense of discomfort as they eagerly hopped in a car to escape the computational gaze.

"Man, it's good to be out of there. They call this a free society when you wind up behind bars for six month for nothing. Walking to pick up my kid at school and next thing you know. How's my daughter doing? Thanks for watching after her."

"Traumatized no doubt. She don't want to be around computers anymore I can tell you that much. Sit a phone on the table and she'll put a napkin over it to cover the cameras. Tells people ”computers” took her Dad."

"Humans took her Dad. Computers told them I was a danger. That's why it took my case six months to get dismissed. The officers were ”working with the information they had” is what I was told."

"You think it would've been different if you stayed calm?"

"I don't know man. You're telling me I have to give in, let the computers oppress us now? I'm three blocks from my daughter's elementary school and I get cut off by a cop car. They ask for my ID and tell me they're gonna pat me down. All because one of those cameras saw my face and decided I was a threat? You better believe I'm losing my temper in that situation."

"So those cameras are broken, is what's up? That's why they let you out? What's that even mean?"

"Yeah man. Algorithmic bias, they call it. Apparently no one warned these programmers that past examples might teach their little computers to see a black face and think criminal."

"Well stick with this class action suit your lawyer got you in on. Take a small chunk of those giant paychecks these technology companies receive."

Ray and his brother went silent as a cop car with a mounted camera crept past them on the highway.

Figure 3. Deeply Learned Bias (Scenario 10)

face recognition, it seems quite likely CV's application to policing will continue to grow. Of course, the goal of the research community should be to diminish bias rather than exacerbate and obfuscate it. One must also understand that, while a prejudice police officer adds harm and reduces trust, this person can be isolated and ideally, punished. If a widespread vision system was found to be biased, the implications to trust could easily fall on an entire industry with the camera acting as a symbol. In an industry that already has systematic disparities in demographic representation [43], care should be taken to ensure that the applications, training sets, and best practices avoid the growth of such a scenario.

Scenario 21 (Figure 4) presents us with a working woman who was impacted by a major attack against the power grid. Seeded by vulnerable security cameras, a worm

"Take all home surveillance systems offline," they were told. Until security experts could come up with a solution, people were forced to power down their networks of devices. Up until last week, the dispersion of networked devices around the home were a boon. Your coffee was ready when you woke up, your thermostat adjusted itself to the erratic weather patterns, and faithful cameras watched over your home, products, and in some case children. Sarah sat in her shadowy home, waiting for the phone to ring for an interview with the New York Times about her experience in the blackout. Power was back up, but she preferred to leave as much equipment off as possible since she didn't really understand which devices could be hacked or not.

"Hi, Sarah? This is Preet Singh with the New York Times. Is now a good time?"

"As good a time as any."

"OK, I'm going to start recording now, please let me know if there's anything you'd like excluded from print. Tell me a bit about what happened when your power first went out."

"Well I was arriving home from work and normally my garage just opens for me when it sees my car or license plate or however it works. But it wouldn't open. So I went on my phone and tried to open it manually and that did nothing either. Only then did it hit me that it must have been a bigger outage since the red light down the block was out too. Being a bright and sunny day, I was very confused and to be honest scared."

"That's understandable. And what can you tell me about what you've learned since then about the situation?"

"Well from what I can tell, I got the least of it out here in the suburbs. It was mayhem in the cities. Now I don't really understand the details, but apparently my home security system, thermostat, everything really, might have participated in the attack. Something happened here for sure. Many fuses in my home were blown out. I guess moments before I arrived home everything surged. Is that right?"

"That's right. What's known is there was a large-scale exploit that started with camera systems connected to the internet. But now that these cameras are connected into integrated homes such as smart thermostats and lighting systems, they were able to create timed surges that targeted certain distribution assets in the power grid."

"Oh my.. So someone used a camera to operate my home?"

"Essentially, yes."

"We're always trying to advance so fast, told to buy the next product. Doesn't anyone test these things first?"

"Of course ma'am, security threats are known, but are very hard to control. No one saw this coming, I can assure you."

"Well I believe that, but city-wide blackouts, my goodness. Whoever thought of this stuff should've warned business before they released so many products."

"Do you believe an event like this will change your future trust in technology products?"

"Absolutely. How could it not? Is all this really worth some minor convenience?"

Figure 4. The Cameras Attack (Scenario 21)

made it to seemingly millions of homes to create timed power surges. This was made possible due to a supposed advancement in integrated or smart homes. The story suggests that many homes have become equipped with intelligent camera systems, thermostats, and appliances, allowing an attack on any one system to be threatening to the whole ecosystem. One assumption is that home network security has not improved significantly in the interval between now and the context of the fictional tale. It is also taken for granted that people continue to be lax about data sharing between devices and systems. Given that a camera system could act as a critical part of my other smart-home devices, we assume the camera would be connected to nearly everything in the home. We also conjecture that an adversary who understands the intricacies of the power grid could also design an adversarial system that could precisely surge power in homes, placing a critical load on particular assets.

How far out is such a scenario? To what extent does it really implicate CV researchers? It should immediately strike readers that leveraging insecure, distributed devices is a reality of our time that is unlikely to go away. On Friday October 21, 2016, we saw the largest DDoS attack ever, using IoT devices, particularly cameras, to deliver 1.2Tbps targeted at DNS provider Dyn [49]. Security experts have warned that these attacks are likely to grow in size and frequency [5], and that the market is not the place to look for solutions [49]. What makes this issue particularly tricky for CV practitioners is that unless the computer power required to perform video processing significantly diminishes, most systems using a video feed will require internet access. That is, much like Google's NLP engine relies on cloud services to process audio samples, CV seems destined for a similar future in the cloud. In the long-run, large-scale attacks could cause both blanket harm to society and mass distrust for using cloud-enabled or IoT devices. A moment such as the one described in this narrative would likely signify the necessity of government intervention on the problem which, if impulsive, could severely deter industry development and limit the applicability of research insights.

5. Conclusion & Future Work

Emerging from this dive into the risky situations that CV research might lead us into, we see a number of takeaways applicable to further develop ethics in this field. To begin with, we have postulated and explored a variety of scenarios where some sort of disparate, yet significant impact is enacted through bias and discrimination. Preventing this reality will take a lot of work, but has tangible ways forward. One way forward could be to seek professional certifications for particular practitioners who design systems where the results have serious life consequences. Much like certifications for doctors, lawyers, architects, and professional engineers, we could see sub-fields of computer science adopt licensure programs. Something that can be done sooner is taking seriously our responsibility to perform blackbox tests, audit our systems, and provide access to unbiased datasets. These efforts already have early starting points with the Algorithmic Justice League and the Fair, Accountable, and Transparent Machine Learning Conference.

Another place we may consider diverting expertise into is both for- and non-profit oversight projects. Much like cybersecurity has pen-testers who work toward bug bounties set out to prevent major hacks, we could imagine adversary bounties where researchers prove the ability to create adversarial examples to systems before they go live. Similarly, we could see public-interest groups who certify particular systems as fair, using their seal of approval to aid the public in choosing systems developed by best practices.

Other mechanisms that may help mitigate some of these risks are working sooner, rather than later, with policy experts to advocate for security standards on IoT devices and routers. The more distance between experts and policymakers, the higher likelihood the eventual policies designed will be damaging to the field. In the same vein, to prevent some of the concerns around privacy and de-anonymization, there are already regulatory models, like the EU's GDRPs, that attempt to give users more control and knowledge over who owns and uses their data. While this could be seen as a short-term inefficiency to data mining operations, it may prevent a long-term turn away from the field as the abundance and severity of privacy harms develop. Last, but certainly not least, is experts weighing in on the kinds of systems that should keep a human in the loop. It is exciting to see the accuracy and capability of CV work grow, but it is critical that practitioners recognize the limits of what sorts of judgements we want automated and where checks and balances should exist.

As a broad effort, this work points easily toward further deep research on the topic of risk in CV. Surveying more experts in the field about risky scenarios and future applications could enrich the assessment and help the public and policymakers understand what emerging trends are most dangerous. Further, given the interrelationship of psychological harm, trust, and technical knowledge, information could be gathered from users of these systems to get a more targeted assessment of how different populations perceive these risks. Finally, CV researchers, and computer scientists at large, should actively determine how they can make ethics a central part of their concentration area. Incentivizing ethical innovation must be a major factor to any serious interest in warding off dangerous or harmful problems in an expert domain. Emphasizing the ethical dimensions of CV research and taking seriously the study of risk factors such as those discussed in this paper will ensure a prosperous and fair future of the field.

References

[1] All Eyes On You.

[2] The Perpetual Line-Up.

[3] Predict Crime | Predictive Policing Software.

[4] Google apologises for Photos app's racist blunder. *BBC News*, July 2015.

[5] New DDoS attack technique could unleash devastating internet meltdown warn experts, Oct. 2016.

[6] J. Anaya and A. Barbu. RENOIR - A Dataset for Real Low-Light Image Noise Reduction. *arXiv:1409.8230 [cs]*, Sept. 2014. arXiv: 1409.8230.

[7] D. Anthony and F. Raby. *Speculative Everything: Design, Fiction, and Social Dreaming*. London, England: The Mit Press, Cambridge, Massachusetts, 2013.

[8] A. M. Barry-Jester. Should Prison Sentences Be Based On Crimes That Havent Been Committed Yet?, Aug. 2015.

[9] C. F. Benitez-Quiroz, R. Srinivasan, Q. Feng, Y. Wang, and A. M. Martinez. EmotioNet Challenge: Recognition of facial expressions of emotion in the wild. *arXiv:1703.01210 [cs]*, Mar. 2017. arXiv: 1703.01210.

[10] B. Bhattarai, G. Sharma, and F. Jurie. CP-mtML: Coupled projection multi-task metric learning for large scale face retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4226–4235, 2016.

[11] J. Bleecker. Design Fiction: A short essay on design, science, fact and fiction. *Near Future Laboratory*, 29, 2009.

[12] M. Blythe. The Hitchhiker's Guide to Ubicomp: Using Techniques from Literary and Critical Theory to Reframe Scientific Agendas. *Personal Ubiquitous Comput.*, 18(4):795–808, Apr. 2014.

[13] M. Blythe. Research Through Design Fiction: Narrative in Real and Imaginary Abstracts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pages 703–712, New York, NY, USA, 2014. ACM.

[14] K. W. Bowyer and P. J. Flynn. The ND-IRIS-0405 Iris Image Dataset. *arXiv:1606.04853 [cs]*, June 2016. arXiv: 1606.04853.

[15] J. Buolamwini. AJL -ALGORITHMIC JUSTICE LEAGUE.

[16] J. Buolamwini. The Algorithmic Justice League, Dec. 2016.

[17] L. Cavigelli, D. Bernath, M. Magno, and L. Benini. Computationally efficient target classification in multispectral image data with Deep Neural Networks. In *SPIE Security+ Defence*, pages 99970L–99970L. International Society for Optics and Photonics, 2016.

[18] C. Chambers. The psychology of mass government surveillance: How do the public respond and is it changing our behaviour? *The Guardian*, Mar. 2015.

[19] Y.-L. Chen, T.-W. Huang, K.-H. Chang, Y.-C. Tsai, H.-T. Chen, and B.-Y. Chen. Quantitative Analysis of Automatic Image Cropping Algorithms: A Dataset and Comparative Study. *arXiv:1701.01480 [cs]*, Jan. 2017. arXiv: 1701.01480.

[20] T. Coughlan, M. Brown, G. Lawson, R. Mortier, R. J. Houghton, and M. Goulden. Tailored Scenarios: A Low-cost Online Method to Elicit Perceptions on Designs Using Real Relationships. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '13, pages 343–348, New York, NY, USA, 2013. ACM.

[21] N. R. Council and others. *Understanding risk: Informing decisions in a democratic society*. National Academies Press, 1996.

[22] K. Crawford and J. Schultz. Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms. SSRN Scholarly Paper ID 2325784, Social Science Research Network, Rochester, NY, Oct. 2013.

[23] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.

[24] C. DiSalvo. Spectacles and Tropes: Speculative Design and Contemporary Food Cultures. *The Fibreculture Journal*, (20 2012: Networked Utopias and Speculative Futures), 2012.

[25] L. Eckhouse. Big data may be reinforcing racial bias in the criminal justice system. *The Washington Post*, Feb. 2017.

[26] B. Fischhoff, S. R. Watson, and C. Hope. Defining risk. *Policy Sciences*, 17(2):123–139, 1984.

[27] A. Hadid. Face biometrics under spoofing attacks: Vulnerabilities, countermeasures, open issues, and research directions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 113–118, 2014.

[28] B. Hasani and M. H. Mahoor. Spatio-Temporal Facial Expression Recognition Using Convolutional Neural Networks and Conditional Random Fields. *arXiv:1703.06995 [cs]*, Mar. 2017. arXiv: 1703.06995.

[29] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. *arXiv:1512.03385 [cs]*, Dec. 2015. arXiv: 1512.03385.

[30] L. A. Hendricks, S. Venugopalan, M. Rohrbach, R. Mooney, K. Saenko, and T. Darrell. Deep Compositional Captioning: Describing Novel Object Categories without Paired Training Data. *arXiv:1511.05284 [cs]*, Nov. 2015. arXiv: 1511.05284.

[31] S. Jasanoff. Technologies of humility: citizen participation in governing science. *Minerva*, 41(3):223–244, 2003.

[32] J. Jouvenal. Police are using software to predict crime. Is it a holy grail or biased against minorities? *Washington Post*, Nov. 2016.

[33] Y. Kim, B. Yoo, Y. Kwak, C. Choi, and J. Kim. Deep generative-contrastive networks for facial expression recognition. *arXiv:1703.07140 [cs]*, Mar. 2017. arXiv: 1703.07140.

[34] J. L. L. J. A. Kirchner, Surya Mattu. Machine Bias: Theres Software Used Across the Country to Predict Future Criminals. And its Biased Against Blacks., May 2016.

[35] E. Kocabey, M. Camurcu, F. Ofli, Y. Aytar, J. Marin, A. Torralba, and I. Weber. Face-to-BMI: Using Computer Vision to Infer Body Mass Index on Social Media. *arXiv:1703.03156 [cs]*, Mar. 2017. arXiv: 1703.03156.

[36] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[37] M. Kumar, A. Veeraraghavan, and A. Sabharwal. DistancePPG: Robust non-contact vital signs monitoring using a camera. *Biomedical optics express*, 6(5):1565–1588, 2015.

[38] A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial examples in the physical world. *arXiv:1607.02533 [cs, stat]*, July 2016. arXiv: 1607.02533.

[39] G. Levi and T. Hassner. Age and gender classification using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 34–42, 2015.

[40] J. Mankoff, J. A. Rode, and H. Faste. Looking Past Yesterday's Tomorrow: Using Futures Studies Methods to Extend the Research Horizon. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 1629–1638, New York, NY, USA, 2013. ACM.

[41] J. Mendoza. 'Predictive policing' isn't in science fiction, it's in Sacramento. *Christian Science Monitor*, Aug. 2016.

[42] A. Nguyen, J. Yosinski, and J. Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 427–436, 2015.

[43] M. Nisen. Only 2% of Google's American Workforce Is Black. *The Atlantic*, May 2014.

[44] W. Odom, J. Zimmerman, S. Davidoff, J. Forlizzi, A. K. Dey, and M. K. Lee. A Fieldwork of the Future with User Enactments. In *Proceedings of the Designing Interactive Systems Conference*, DIS '12, pages 338–347, New York, NY, USA, 2012. ACM.

[45] S. J. Oh, R. Benenson, M. Fritz, and B. Schiele. Faceless person recognition: Privacy implications in social media. In *European Conference on Computer Vision*, pages 19–35. Springer, 2016.

[46] F. Pasquale. *The black box society: The secret algorithms that control money and information*. Harvard University Press, 2015.

[47] T. Poggio and Q. Liao. Theory II: Landscape of the Empirical Risk in Deep Learning. *arXiv:1703.09833 [cs]*, Mar. 2017. arXiv: 1703.09833.

[48] R. Ramirez, M. Mukherjee, S. Vezzoli, and A. M. Kramer. Scenarios as a scholarly methodology to produce interesting research. *Futures*, 71:70–87, Aug. 2015.

[49] B. Schneier. Essays: Lessons From the Dyn DDoS Attack - Schneier on Security.

[50] K. Sikka, G. Sharma, and M. Bartlett. Lomo: Latent ordinal model for facial analysis in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5580–5589, 2016.

[51] J. Stanley. Does Surveillance Affect Us Even When We Cant Confirm Were Being Watched? Lessons From Behind the Iron Curtain.

[52] B. Sterling. Design Fiction. *interactions*, 16(3):20–24, May 2009.

[53] S. Thielman. Can we secure the internet of things in time to prevent another cyber-attack? *The Guardian*, Oct. 2016.

[54] C. Timberg. Racial profiling, by a computer? Police facial-ID tech raises civil rights concerns., Oct. 2016.

[55] L. Vassie, P. Slovic, B. Fischhoff, and S. Lichtenstein. Facts and fears: understanding perceived risk. *Policy and Practice in Health and Safety*, 3(sup1):65–102, 2005.

[56] J. Vincent. Magic AI: these are the optical illusions that trick, fool, and flummox computers, Apr. 2017.

[57] J. Voros. A primer on futures studies, foresight and the use of scenarios. *Prospect: The Foresight Bulletin*, 6(1), 2001.

[58] W. Wei. Two Arrested for Hacking Washington CCTV Cameras Before Trump Inauguration.

[59] L. Winner. Do artifacts have politics? *Daedalus*, pages 121–136, 1980.

[60] X. Wu and X. Zhang. Automated Inference on Criminality using Face Images. *arXiv:1611.04135 [cs]*, Nov. 2016. arXiv: 1611.04135.

[61] X. Wu, X. Zhang, and C. Liu. Automated Inference on Sociopsychological Impressions of Attractive Female Faces. *arXiv:1612.04158 [cs]*, Dec. 2016. arXiv: 1612.04158.

[62] C. Xie, J. Wang, Z. Zhang, Y. Zhou, L. Xie, and A. Yuille. Adversarial Examples for Semantic Segmentation and Object Detection. *arXiv:1703.08603 [cs]*, Mar. 2017. arXiv: 1703.08603.

[63] Y. Xu, J. Heinly, A. M. White, F. Monrose, and J.-M. Frahm. Seeing double: Reconstructing obscured typed input from repeated compromising reflections. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*, pages 1063–1074. ACM, 2013.

[64] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. Stacked Attention Networks for Image Question Answering. *arXiv:1511.02274 [cs]*, Nov. 2015. arXiv: 1511.02274.

[65] K. Zetter. An Unprecedented Look at Stuxnet, the Worlds First Digital Weapon. *WIRED*, Nov. 2014.

[66] S. Zhang, G. Wu, J. P. Costeira, and J. M. F. Moura. Understanding Traffic Density from Large-Scale Web Camera Data. *arXiv:1703.05868 [cs]*, Mar. 2017. arXiv: 1703.05868.