BUG BOUNTIES FOR ALGORITHMIC HARMS?

Lessons from cybersecurity vulnerability disclosure for algorithmic harms discovery, disclosure, and redress

December 2021

ALGORITHMIC JUSTICE LEAGUE

BUG BOUNTIES FOR ALGORITHMIC HARMS?

Lessons from cybersecurity vulnerability disclosure for algorithmic harms discovery, disclosure, and redress

Josh Kenway and Camille François
Algorithmic Justice League

December 2021

Authored by Josh Kenway and Camille François.

Edited by Dr. Sasha Costanza-Chock.

Guidance and revisions by Dr. Joy Buolamwini and Inioluwa Deborah Raji.

Cover art and illustrations by clarote (clarote.net).

Report design and layout by Casa Blue (casablue.com).

Produced and published by the Algorithmic Justice League (ajl.org).

This report was made possible in part by funding from the Alfred P. Sloan Foundation, the Rockefeller Foundation, and the Mozilla Foundation.

ALGORITHMIC JUSTICE LEAGUE

Suggested Citation
Kenway, Josh, and Camille François. Bug Bounties For Algorithmic Harms?
Lessons from Cybersecurity Vulnerability Disclosure for Algorithmic Harms Discovery, Disclosure, and Redress.
Edited by Sasha Costanza-Chock. Washington, DC: Justice League. December 2021.
Available at https://ajl.org/bugs.

ACKNOWLEDGMENTS

The authors acknowledge the invaluable contributions of other members of the Algorithmic Justice League and the CRASH team to this project. In particular, we are grateful to Dr. Sasha Costanza-Chock for her extensive edits and feedback on this report, and Inioluwa Deborah Raji and Dr. Joy Buolamwini for their detailed revisions and many other integral contributions to this work.

We conducted a series of interviews with the following experts and practitioners, who very generously shared their time and experiences with us: Alex Rice, Amit Elazari Bar On, Dino Dai Zovi, Jack Cable, Lisa Wiswell Coe, Marcia Hofmann, Mårten Mickos, and Rayna Stamboliyska. Summaries of their interviews can be found here.

We are also indebted to our many colleagues who helped us navigate the scattered and varied body of relevant works across academic papers, corporate and news media, hacker forums, books, program materials, blog posts, infosec talks, and so on. We would like to offer our thanks to Matt Goerzen, Yuan Stevens, and Ryan Ellis for their crucial role in helping us work through aspects of the history of BBPs, as well as fundamental considerations related to power, community, and market dynamics at play within the BBP ecosystem. We also thank Dana Tzegaegbe and Akoth Ombaka for their operational support in driving this research project forward, as well as other members and friends of AJL and the CRASH project for helping to shape our thinking on the various topics encompassed in this report.

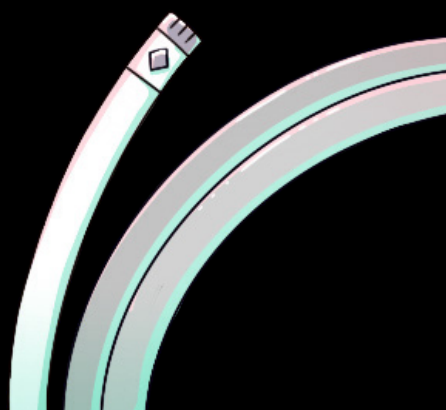
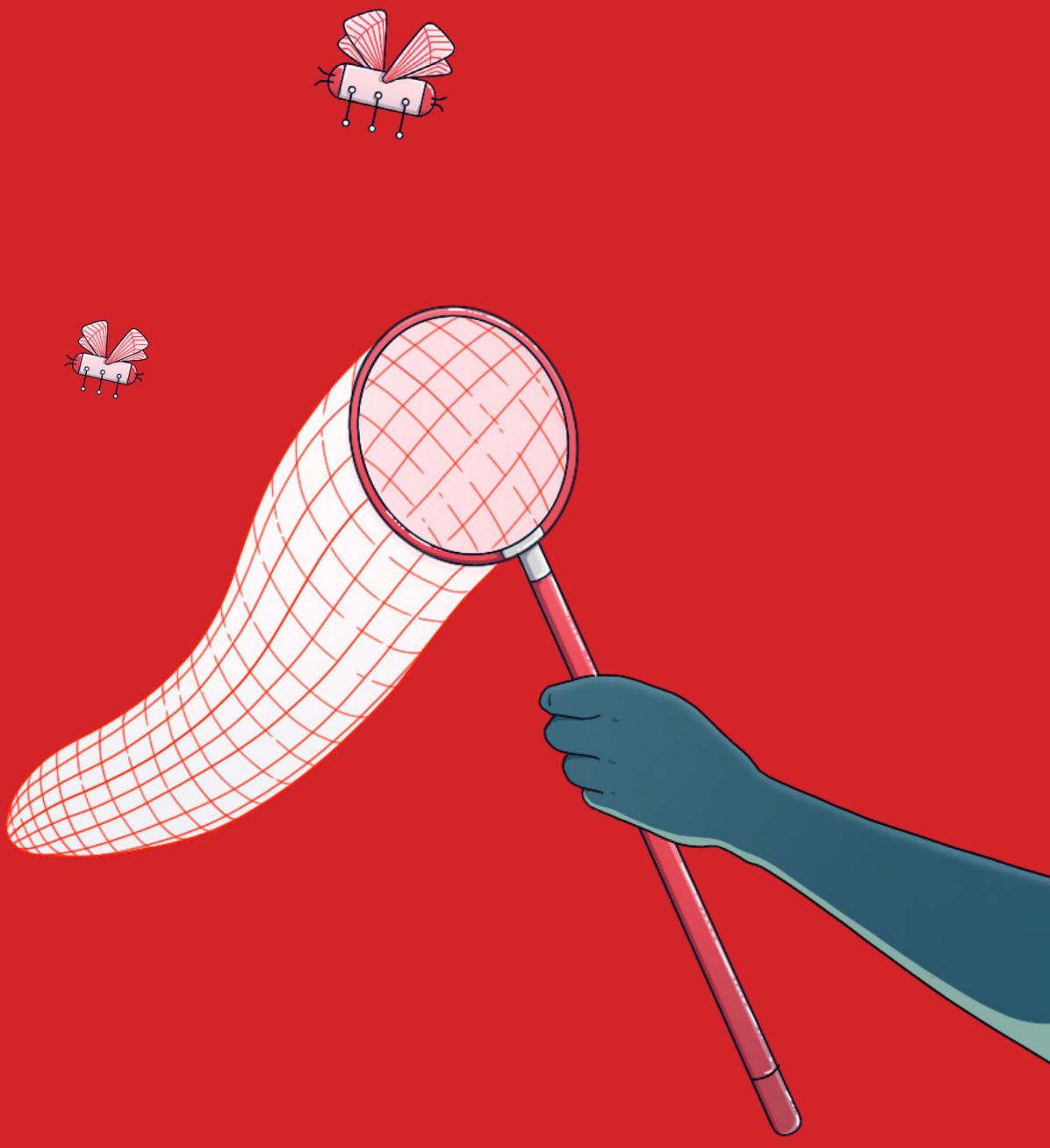


TABLE OF CONTENTS

EXECUTIVE SUMMARY



Paying hackers to disclose bugs was once considered radical; now, it's common. 'Bug bounty[1]' programs (BBPs) for cybersecurity vulnerabilities, wherein participants are rewarded for identifying exploitable flaws (or security 'bugs') in software or hardware, are increasingly popular. Google, the Department of Defense, Starbucks, and hundreds of other companies and organizations regularly use BBPs to buy security flaws from hackers. A wide variety of organizations have adopted BBPs, and a growing number of people participate, most often via a small number of platforms that have come to be the preferred host of such programs (such as HackerOne or BugCrowd). BBPs can be understood within a broader context of third-party vulnerability research and reporting mechanisms, such as bounty-less vulnerability disclosure programs (VDPs) and contract- or employment-based penetration testing ('pentesting'). BBPs, VDPs, and pentesting have also been adopted to address a wider spectrum of socio-technical harms and risks beyond security bugs. For example, programs from Facebook, Twitter, and others have recently begun to extend the BBP model to address an expanded set of socio-technical harms. Yet the conditions under which BBPs might constitute appropriate mechanisms for addressing socio-technical concerns remain relatively unexamined. [2]

The Algorithmic Justice League (AJL)'s Community Reporting of Algorithmic System Harms (CRASH) project is focused on creating an inclusive community of researchers, practitioners, and everyday people who have been harmed by algorithmic systems. We believe that together we can take action to help prevent, identify, report, and ensure redress for algorithmic harms. Our exploration of whether bounties might usefully be applied to algorithmic harms discovery was motivated both by the promises of such systems and by an awareness of the 'buzz' around BBPs, reflected in the proliferation of security BBPs and in the nascent adoption of BBPs for a wider variety of socio-technical issues.

In the Introduction to this report, we share our objectives and methods. We describe how we set out to explore the overall merits and design lessons of BBPs, as well as closely related vulnerability discovery and disclosure models. The report focuses on the applicability of bounties to the algorithmic harms context, specifically, as well as to socio-technical concerns more broadly. Our methods include interviews with BBP experts and practitioners, review of existing literature, and analysis of historical and present-day approaches to vulnerability disclosure, in order to arrive at a set of key findings and concrete recommendations. Specifically, we explore three broad lines of inquiry related to how BBPs might be used to:

FOSTER AND NURTURE PARTICIPATION AND COMMUNITY AMONG RESEARCHERS

SHAPE FIELD DEVELOPMENT BY FOSTERING THE DEVELOPMENT OF RESOURCES AND METHODS

DRIVE TRANSPARENCY AND ACCOUNTABILITY ACROSS THE INDUSTRY

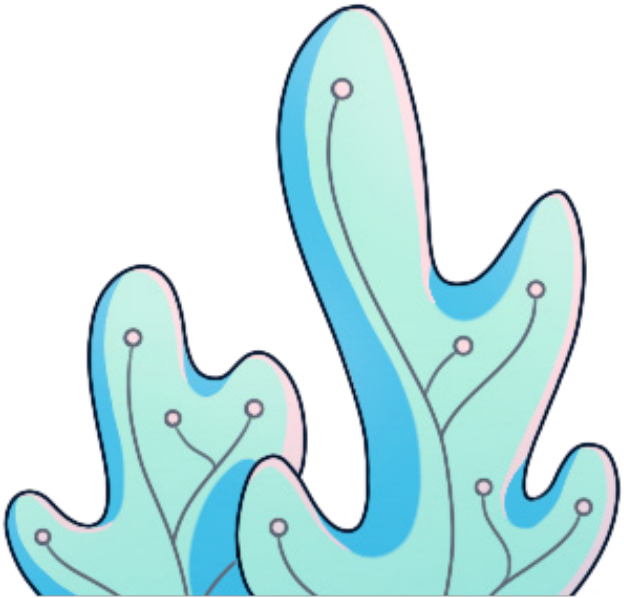
1 We use the word 'bounty' throughout this report as it is widely used throughout the industry to describe these programs. However, we also recognize the term's problematic history of use by those who perpetuated slavery as they sought to find and re-enslave those who escaped conditions of slavery.

2 Important preliminary work on these questions include Elazari (2018), who explores legal considerations around BBPs for flawed algorithms, and Ellis and Stevens (forthcoming) on the labor ethics of cybersecurity BBPs.

In section II. From Cybersecurity Vulnerabilities to Algorithmic Harms, we start with key definitions, then compare cybersecurity vulnerabilities and algorithmic harms along both conceptual and practical lines. In section III. Bug Bounty Programs 101, we provide historical background for understanding the emergence of BBPs and related vulnerability disclosure mechanisms, including the platforms that often facilitate them. We develop and share a design framework for vulnerability disclosure mechanisms, provide a taxonomy of typical vulnerability reporting mechanisms with reference examples, and identify the circumstances under which various design configurations might be useful for surfacing and addressing harmful flaws. Our design levers — captured in full in Figure 1: Design Levers Across the BBP Spectrum — differentiate program design along the following dimensions:

- **Target Entities** — Does a particular program or platform solicit reports only for vulnerabilities affecting organizations that seek to receive such reports, or can reports be submitted even for non-participating organizations?

- **Compensation Model** — How are security researchers compensated for their contributions and expertise?



- **Disclosure Model** — Under what terms are researchers who discover a vulnerability authorized to publicly disclose their findings (e.g., to the press, in blog posts, or academic research), and on what timeline?

- **Participation Model** — To what extent are mechanisms intended for widespread public participation ("open") versus permitting only a numerically limited number of select researchers ("closed")?

- **Program Management** — To what extent are the responsibilities of program management handled directly by the organization that seeks to receive reports, versus by third-party platforms? (For example, hosting program terms, receiving reports, validating submissions, triaging vulnerabilities, facilitating patch development, and verifying patches).

- **Program Duration** — Are programs intended to be temporary, or long-lived?

- **Program Scope & Access** — How wide is the range of systems and vulnerability types formally encompassed under a given program, and what level of technical and organizational access is afforded to researchers to facilitate exploration and analysis?

In section IV. Key Research Takeaways, we synthesize five high-level findings that can be summarized as follows, alongside relevant recommendations that emerged through our analysis of design levers and lessons:

KEY TAKEAWAY

1. PREPARE TO INCLUDE SOCIO-TECHNICAL

A handful of players in the BBP ecosystem have been slowly expanding their current programs to include socio-technical issues.

For example, BBPs have been set up to identify instances of data abuse (Google and Facebook), systematic errors with video game cheat-flagging algorithms (Rockstar Games), bias in image cropping algorithms for social feeds (Twitter), and deficient privacy protections in mobile software relative to vendors' claims (Correlium). This progression hasn't happened in a structured way, and no clear best practices have emerged yet, but the trend is likely to continue accelerating.

RECOMMENDATIONS

- Carefully consider what kinds of socio-technical issues (e.g., data abuse, algorithmic harm, platform abuse, and so on) a BBP could most usefully help to unearth for their organization, as well as whether the organization is ready to commit the resources needed to ensure that any problems are addressed.

- Review relevant design levers and findings before setting up BBPs in order to avoid repeating known bad practices from the cybersecurity domain (refer to Figure 1: Design Levers Across the BBP Spectrum and Design Companion: Lessons from BBPs).

- Be aware that reporting templates can shape the priorities and direction of the field. Thoughtfully designed reporting templates and similar tooling for surfacing and reporting algorithmic harms will be indispensable not only for BBPs but more widely for understanding and addressing these issues.

"[IF] AN INFORMATION SECURITY TEAM LAUNCHES A BOUNTY PROGRAM, THEY'RE ONLY GOING TO INCENTIVIZE THE THINGS THAT THEY'RE IN A POSITION TO DO SOMETHING ABOUT … [IT'S] NOT LIKE 'ACME INC.' LAUNCHES A BOUNTY PROGRAM; YOU HAVE TO PEEL IT BACK … TO [ASK] WHICH TEAM LAUNCHED THAT PROGRAM [AND] WHAT TYPE OF FEEDBACK ARE THEY IN A POSITION TO DO SOMETHING ABOUT?"

— Alex Rice, CTO, HackerOne[3]

3 AJL Interview: Alex Rice

KEY TAKEAWAY

2. LOOK ACROSS THE LIFECYCLE.

The allure of BBPs can often obscure the fact that they are just one mechanism for enhancing cybersecurity that fits within a broader, ongoing arc of field maturation and organizational development.

A similar 'technology product lifecycle' lens to that which exists in security (as the 'secure development lifecycle') should be applied to algorithmic harms, since to be responsive to reports of algorithmic harm, organizations will need to develop the 'digestive systems' to prioritize, assess, and act on those findings. [4]

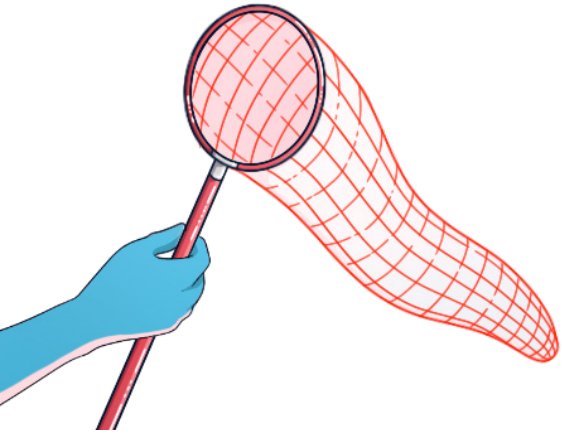
RECOMMENDATIONS

- To maximize the value of a BBP-like approach, organizations should establish approaches to addressing algorithmic harms across the full product lifecycle, and commit the resources to address discovered problems at the root.

- Researchers focused on algorithmic bias and harm have often emphasized the need to consider harms beyond a narrow focus on input data or model specification. Practitioners may look to 'secure development lifecycle' methodologies for inspiration.

"[THE] BEST BUG BOUNTY CUSTOMERS WE HAVE, THEY TAKE EVERY BUG, THEY GO BACK TO SOFTWARE DEVELOPMENT AND SAY, 'WHAT CAN WE LEARN FROM THIS BUG ON A DEEPER LEVEL?' NOT JUST FIXING [THAT VULNERABILITY], BUT FIXING THE WAY [THEY] DEVELOP SOFTWARE."

— Mårten Mickos, CEO, HackerOne[5]



4 Relatedly, scoring methodologies for algorithmic harms bounties will need to be context-sensitive to a greater extent than has so far been the case for cybersecurity vulnerabilities.

5 AJL Interview: Mårten Mickos

KEY TAKEAWAY

3. NURTURE THE COMMUNITY OF PRACTICE.

Bug bounty platforms play both a direct and indirect role in nurturing communities of practice. For example, they may provide educational materials and tools, and motivate community members to independently develop and share resources.

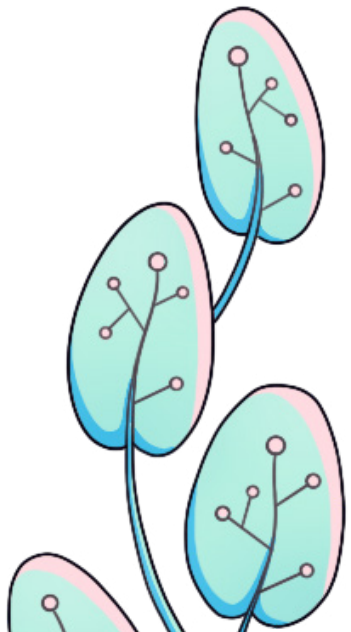
We believe there is a strong need for similar, well-curated, accessible resources to help nurture the algorithmic harms research community. We caution against community-building approaches that exclude researchers from fields outside of computer science and community advocates while unduly elevating technical skills and techno-solutionism.

RECOMMENDATIONS

- Bounty program and platform affordances and legal terms should actively encourage collaboration and ensure fair compensation.
- Programs and platforms can help foster a community of practice through education, information sharing, and trust-building, for example, through the development and consolidation of formal learning tools and resources, publicization of past reports, and tracking of community standing / reputation based upon useful contributions.
- Practitioners inside organizations that produce or use algorithmic systems should look externally for how they can engage and support the development of the wider community.

"[IF] RESEARCHERS REALLY DO START TO FEEL A PART OF THE BROADER TEAM … THAT'S HOW YOU'RE GOING TO GET [THAT] GROUP … TO CONTINUE TO WANT TO DO THIS FOR THE LONG RUN."

— Lisa Wiswell Coe, Fmr. Program Manager, Hack the Pentagon[6]



6 AJL Interview: Lisa Wiswell Coe

KEY TAKEAWAY

4. INTENTIONALLY DEVELOP A DIVERSE, INCLUSIVE COMMUNITY.

A small number of bug bounty platforms dominate the vulnerability disclosure ecosystem, leading to concerns about the commodification of security research and lack of diversity among these platforms' contributors.

Many vulnerability discovery mechanisms can be understood as a form of outsourcing, with compensation arrangements ranging from nothing to uncertain to incomplete, with few paths to stable employment. The fact that a handful of highly-skilled bug hunters are able to make a comfortable living from BBPs should also not obscure the fact that the vast majority of bug hunters make very little, and that these programs tend to outsource security work in bulk, while doing little to substantively shift longstanding demographic inequalities within the cybersecurity field. Deploying BBPs successfully for algorithmic harms will require serious effort to recruit and retain diverse communities of researchers and community advocates, and ensure fair compensation for work.

RECOMMENDATIONS

- Platforms and programs should aim to cultivate diverse and inclusive communities, and set meaningful, time-bound targets for those objectives, including to ensure that those with lived experience of the impacts of harmful systems are included, as well as both technical and socio-technical experts.

- Practitioners, advocates, and others across the ecosystem should prioritize inclusive and participatory processes to establish standards, frameworks, templates, and tooling.

"THE PLATFORMS ARE TRYING TO EXPAND THEIR COMMUNITIES … [AND] EDUCATE PEOPLE AND BROADEN THE SCOPE OF WHO'S PARTICIPATING … BUT EVEN THEN … IT REMAINS THAT THERE'S A RELATIVELY SMALL NUMBER OF PEOPLE WHO FIND MOST OF THE VULNERABILITIES [AND] THEY'RE NOT VERY DIVERSE."

— Jack Cable, Security Researcher[7]

7 AJL Interview: Jack Cable

KEY TAKEAWAY

5. FOSTER AND PROTECT PARTICIPATORY, ADVERSARIAL RESEARCH, AND GUARANTEE SOME FORM OF PUBLIC DISCLOSURE.

Vulnerability disclosure continues to struggle with models for compensated, safe, third-party or adversarial research.

Platforms that operate BBPs and act as intermediaries between hackers and target organizations play an outsized role in the current BBP ecosystem, with target organizations in turn typically afforded a right of non-disclosure. The business models and funding sources of BBP platforms tend to bend their decision-making towards the interests of target organizations, for example, in determining whether vulnerability reports can be publicly released. Greater protection for third-party algorithmic harms research is sorely needed.

RECOMMENDATIONS

- Platforms must take care to avoid capture by vendors and/or operators. Uncritically importing current BBP structures into socio-technical fields may actually lead to less disclosure and more suppression of critical research.
- Researchers who investigate and disclose algorithmic harms face complex legal risks. Those who organize third party research should consider including clear communication about risks, provision of pro-bono support, and advocacy for stronger protections under the law.
- Mechanisms such as legal safe harbor, when offered by target organizations without unfair conditions attached, may also help reduce real and perceived legal risk to contributors.
- Bounty platforms should be designed to facilitate transparency and accountability, even where vendors or operators may prefer to avoid scrutiny.

"[T]HE COMPANY … HAS A LOT OF POWER TO DEFINE THE TERMS OF THE ENGAGEMENT … THEY'RE THE ONES WHO HOLD THE PURSE STRINGS; THEY DECIDE WHEN CONDITIONS ARE MET, AND WHAT THE PAYOUT IS GOING TO BE … I THINK THAT ALL THE PLAYERS IN THE ECOSYSTEM [ARE] … INCENTIVIZED TO MEET THE[IR] REQUIREMENTS."

— Marcia Hofmann, Digital Rights Lawyer[8]

8 AJL Interview: Marcia Hofmann

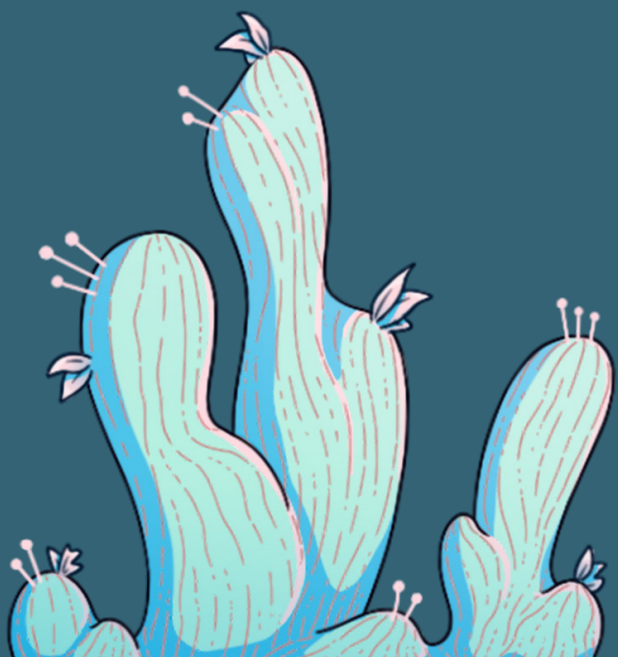
After discussing our key takeaways, in section V. Case Study: Twitter's Algorithmic Bias Bounty Challenge, we then explore Twitter's recent 'bias bounty' challenge as an example of the opportunities and pitfalls of a BBP-like model for algorithmic harms. We evaluate Twitter's first algorithmic bias bounty program, which offered researchers cash rewards for demonstrating bias in Twitter's image salience and cropping algorithms, using our Design Levers framework. We suggest that this program was a 'safe' initial foray for the company, and question the scalability and repeatability of the program in that context. We also identify specific ways that the program's terms and conditions of participation, as well as the scoring methodology, might have been improved. We suggest that certain pitfalls might have been avoided through greater transparency and collaborative engagement with the field ahead of time.

In section VI. Conclusion, we note that the value proposition of directly transposing BBPs from cybersecurity into other domains is highly context-dependent. BBPs to identify flaws in algorithmic systems may be effective under certain conditions, but our work to understand the limitations of the BBP model in the cybersecurity space, and to explore the comparability of cybersecurity vulnerabilities and algorithmic harms, suggests that the uncritical deployment of BBPs in this emerging field is unlikely, on its own, to increase accountability for the design, development, and deployment of algorithmic systems. The hype around BBPs does not always lend itself to thoughtful deployment within a wider, lifecycle-oriented strategy. Absent sufficient internal preparedness and the correct alignment of institutional incentives towards addressing vulnerabilities as they are identified, BBPs are rendered little more than an act of security theater. To meaningfully benefit from BBPs and similar vulnerability discovery mechanisms, organizations must already be deeply invested in understanding, addressing, and preventing the problems that second- and third-party researchers uncover.

Despite these many challenges, existing BBPs and related disclosure mechanisms, as well as the platforms that host them, present a number of constructive design lessons for the development of mechanisms for the discovery and disclosure of socio-technical issues, including algorithmic bias and harms. In a Design Companion appended to the main body of the report, we offer 25 design lessons from BBPs that we hope can help inform algorithmic harms reporting work in the future. In the Interview Summaries section, we provide summaries of interviews with key practitioners and experts that we conducted for the report. We also provide a Glossary and a Bibliography.

We hope that our work provides a useful touchstone for those interested in creating participatory mechanisms for the discovery, disclosure, and redress of cybersecurity flaws, algorithmic bias and harms, and other issues in socio-technical systems. Finally, we urge interested readers to connect with the Community Reporting of Algorithmic System Harms (CRASH) Project at crash.ajl.org. Together, let's build a movement for more equitable and accountable AI!

I.
INTRODUCTION



The Algorithmic Justice League (AJL) is on a mission to raise awareness about the impacts of AI, equip advocates with empirical research, build the voice and choice of the most impacted communities, and galvanize researchers, policy makers, and industry practitioners to mitigate algorithmic harms and biases.

In 2020, AJL started the Community Reporting of Algorithmic System Harms (CRASH) project (co-led by Joy Buolamwini, Sasha Costanza-Chock, and Camille François) to bring together key stakeholders for discovery, scoping, and iterative prototyping of tools to enable broader participation in the creation of more accountable, more equitable, and less harmful algorithmic systems. CRASH builds upon the AJL's prior experiments in encouraging people to share their experiences of algorithmic harm with the organization, including through our "Bias in the Wild" initiative, where we crowdsourced reports of harm that people had personally experienced due to biased algorithmic systems.

OBJECTIVES

As a part of the CRASH project, we set out to explore the viability of creating a reporting platform for algorithmic harms similar to bug bounty programs (BBPs) that exist for incentivizing independent security researchers to identify and report security vulnerabilities (known as 'bugs') to be fixed by the relevant system vendor or operator. 'Bounties' can be reputational (halls of fame, reputation metrics, public credit) and / or material (cash bounties, merchandise, tickets, coupons).[9] There is no existing, robust study of BBPs' applicability to other domains in general, or to algorithmic (or 'AI') systems in particular. We therefore decided to conduct this study as necessary due diligence prior to applying the BBP model to other issues, such as algorithmic harms.

Our exploration, and resulting report, comes as BBPs are sometimes framed as a silver bullet for different types of complex issues with, at times, little consideration of their different design levers and associated trade-offs. Our work seeks to go beyond this "BOUNTY EVERYTHING!" meme and explore what, if anything, can be learned from current BBPs to address socio-technical harms, including those that arise through the development or use of algorithmic systems. We hope that this report can be useful for multiple audiences, including practitioners, advocates, and impacted parties seeking to address algorithmic harms,[10] as well as those interested in the evolution of the field of cybersecurity to encompass a wider variety of socio-technical harms (such as information operations, privacy, and more).

METHODS

Our overarching goals led us to adopt a broad interpretation of BBPs and the platforms that host them, in order to best survey principles, lessons, and mechanisms that may be of interest to AJL's mission. Our research examines several aspects of vulnerability disclosure programs (VDPs) and penetration testing programs, which are related, but distinct, from more traditionally defined BBPs.

This work was driven through three distinct lines of effort: a literature review of these topics; interviews

9 Core definitions used in this work are listed in the Glossary.

10 Algorithmic harms are further explored in From Cybersecurity Vulnerabilities to Algorithmic Harms.

with a small number of individuals whom we identified as being well-positioned to speak to the cross-domain considerations of our project; and desk research exploring the existing BBP ecosystem from structural and programmatic perspectives. The literature review included quantitative and qualitative academic papers, news media, industry documentation and best practices, hacker forums, books, program materials, blog posts, infosec talks, and more (a full list of sources can be found in the Bibliography). We conducted a series of interviews with the following experts and practitioners: Alex Rice, Amit Elazari Bar On, Dino Dai Zovi, Jack Cable, Lisa Wiswell Coe, Marcia Hofmann, Mårten Mickos, and Rayna Stamboliyska. Summaries of these interviews are available in the Interview Summaries section.

In exploring what could be learned from the cybersecurity field to help guide the development of mechanisms for participatory disclosure of algorithmic harms, we first considered both the parallels and differences between cybersecurity vulnerabilities and algorithmic harms. While the differences between the two concepts, and by extension between the fields of practitioners who are concerned with identifying and addressing them, are numerous and significant (refer to From Cybersecurity Vulnerabilities to Algorithmic Harms), we found many lessons that are applicable to both fields.

For example, consider the confluence of algorithmic harms and security research in recent scrutiny of algorithmic harms inflicted by Proctorio, a remote proctoring platform that allows educators to surveil students during remote assessments, as well as the response that this criticism elicited from the company (Proctorio, n.d.). Insights cultivated through the experiential expertise of education practitioners, including Canadian education technologist Ian Linkletter, have illuminated the extent to which the service inherently degrades the privacy of students (Feathers 2021). At the same time, technical research has demonstrated that Proctorio's facial recognition technology is both generally ineffective and disproportionately poor at identifying the faces of Black students (Satheesan 2021). Further research has also surfaced concerns that the robustness of security controls needed to protect the confidentiality of product and business data fall woefully short of industry best practices (Satheesan 2020a; 2020b). The vendor's reaction to Linkletter's criticism was to try and intimidate him into silence through legal action (Feathers 2021), alleging in a civil claim that Linkletter "caused [...] copyrighted, confidential and proprietary information to be disseminated online" in violation of the Canadian *Copyright Act* (Supreme Court of British Columbia 2020), while simultaneously dismissing concerns around the efficacy of its facial recognition technology in a letter to U.S. senators (Olsen 2021). This combination of intimidation and recalcitrance in the face of expert scrutiny parallels historic and, to a lesser extent, ongoing dynamics around vulnerability disclosure in the cybersecurity context. At the same time, the use of traditional security research techniques to illuminate the risk of algorithmic harm provides a real-world example of just how proximate these domains can be in practice.

In addition to further exploring the commonality and distinctions between these domains, this report provides a comparative overview of BBPs and related

cybersecurity vulnerability disclosure mechanisms, offers five key takeaways about the existing BBP ecosystem, and briefly analyzes Twitter's 'bias bounty' challenge, launched in mid-2021, through the lens of our design framework and research findings. In a Design Companion appended to the main body of the report, we also offer 25 design lessons from BBPs that we hope can help inform algorithmic harms reporting work moving forward.

Our goal throughout this research project was to develop means for systematically organizing and analyzing discovery and reporting processes from cybersecurity, which have over recent years helped to reduce risks to researchers and increase the overall volume and impact of vulnerability reporting. Leveraging these insights, we have worked to understand the promise and limitations of expanding these mechanisms beyond the security community to support the burgeoning and diverse community of algorithmic harms investigators. We focused on three high-level questions throughout:

(1) **How do BBPs and related cybersecurity vulnerability disclosure mechanisms contribute to or undermine community building?** *These programs emerged out of and are dependent on a community of security researchers. By considering the ways in which they create opportunities for collaboration and exploration, versus impeding such cooperation, our research team sought to learn from these programs about how we might structure mechanisms to support the development of a truly diverse and inclusive community focused on addressing algorithmic harm, from weekend enthusiasts to the next generation of expert researchers.*

(2) **How do these programs advance or constrain the state of knowledge and practice in their field?** *BBPs are associated with the creation of learning materials, the development of relevant tooling, and the maturation of security practices across organizations that receive vulnerability reports. We asked what could be learned to further advance the state of the algorithmic harms research field, including how to make fundamental methods and tools accessible to a greater number of researchers and community advocates.*

(3) **How do these programs enhance or impede transparency and accountability?** *Questions around transparency and accountability have played a central role in the development of vulnerability disclosure practices, and still loom large in the BBP space today. We set out to understand known obstacles to transparency and accountability in the cybersecurity context to provide lessons for the algorithmic harms space, where researchers who expose bias and harms have often been met with adversarial reactions from algorithmic system vendors and operators.*

Additionally, we attempted to approach this work with an awareness of the Design Justice Network Principles in mind (Design Justice Network 2018; Costanza-Chock 2020). Doing so helped us think through which lessons may apply to the algorithmic harms space, and critically engage with ways of improving existing cybersecurity vulnerability reporting mechanisms.

II.
FROM CYBERSECURITY VULNERABILITIES TO ALGORITHMIC HARMS

One of our high level goals was to explore how lessons from cybersecurity might inform the development of algorithmic bias and harms discovery, disclosure, and redress. Our team brought different backgrounds and perspectives to bear in this project, leading to important framing considerations on what is common and what is fundamentally different between cybersecurity vulnerabilities and algorithmic harms. These considerations, which cut across both how we understand these concepts and how they can be addressed, are outlined below, and guide our assessment of which lessons extracted from our examination of BBPs are relevant to the algorithmic harms space.

"[EVEN] W]HEN YOU HAVE … A SECURITY VULNERABILITY, IT ISN'T NECESSARILY CUT AND DRY. YOU STILL HAVE TO GO THROUGH THE PROCESS OF [ASSESSING] WHAT IS [ITS] ACTUAL IMPACT … [AND] YOU HAVE TO MANUALLY ASSESS THAT. I THINK THAT YOU CAN SEE PARALLELS WITH ALGORITHMIC HARMS, OR PRIVACY RELATED FLAWS, WHERE IT'S NOT … NECESSARILY SOMETHING CONCRETE THAT YOU CAN JUST SPELL OUT BUT RATHER, [REQUIRES] MORE OF A PROCESS OF THINKING THROUGH."

— Jack Cable, Security Researcher[12]

DEFINING CYBERSECURITY VULNERABILITIES AND ALGORITHMIC HARMS

For the purpose of this work, we define **cybersecurity**[11] **vulnerabilities** in a manner that aligns with how they are typically understood by cybersecurity practitioners. We also espouse AJL's working definition of algorithmic harms. Both of these definitions can be found in the Glossary.

At the definitional level, vulnerabilities are usually framed within a purportedly objective, overwhelmingly technical, and largely binary construct that implies their presence or eradication as rendering systems either "vulnerable" or "secure." Of course, this notional binary diverges from the reality of cybersecurity as both a highly changeable and inherently socio-technical construct (Goerzen, Watkins, and Lim 2019), given persistent interaction among human and technological elements, including in the introduction, discovery, and management of security vulnerabilities. However, while more holistic risk- and resilience-based approaches to cybersecurity are now commonplace, the field's overall emphasis remains heavily technical, with definitions of "cybersecurity" most often focused on protecting infrastructure and information, rather than the end-users of particular systems,

11 We opt for "cybersecurity" as the field of reference for various concepts discussed in this paper as it is the frame most adopted by the practitioners that we interviewed. However, it is worth noting that this field is also commonly referred to as "infosec" (information security) or computer security. For a thorough examination of these variations in framing the field and their underpinning implications, refer to Nissenbaum (2005), Von Solms and Van Niekerk (2013), Von Solms and Von Solms (2018), and Christen, Gordijn, and Loi (2020).

12 AJL Interview: Jack Cable

let alone the totality of groups and individuals whose security is ultimately impacted (Craigen, Diakun-Thibault, and Purse 2014). This tendency is notably reflected in how cybersecurity practitioners' understanding of vulnerabilities has remained largely unchanged, despite a growing awareness of the socio-technical characteristics of their origins and impact, and broader cognizance of human factors within cybersecurity (Evans et al. 2016). Ultimately, though, vulnerabilities — either as technical or socio-technical constructs — are only ever a factor or mechanism in the causation of cybersecurity harms, since no harm occurs unless the vulnerability is exploited.

"THERE IS AN OVERARCHING TOPIC HERE, WHICH IS, HOW DO WE AS A SOCIETY KEEP EVERYONE … SAFE AND [MAKE TECHNOLOGIES] SAFER? BECAUSE VULNERABLE-BY-DEFAULT IS WHAT REALITIES DICTATE [AND] AS LONG AS VULNERABILITY IS PERMANENT, RESPONSIBILITY AND ACCOUNTABILITY [ALSO NEED TO BE] PERMANENT."

— Rayna Stamboliyska, Fmr. VP Governance and Public Affairs, YesWeHack[13]

In contrast, algorithmic harms are themselves (highly differentiated) outcomes. They can only be further abstracted into measurable impacts "through highly contested and widely variable governance practices, technical requirements, regulations, and documentation" (Metcalf et al. 2021, 736). According to Metcalf et al., rendering algorithmic harms as impacts within the context of accountability mechanisms implicates the social construction of impacts as evaluative objects such that subjective, political judgments are inherent (Ibid., 737). Impacts, as evaluative constructs, "must constantly be scrutinized to ensure that they are adequate and appropriate proxies" for the real-world nature and severity of algorithmic harms (Ibid., 737). Moreover, while efforts to address cybersecurity vulnerabilities tend to focus on providing technical fixes for flaws understood through that same lens, tackling algorithmic harms more obviously necessitates emphasizing real-world harm as the primary outcome to be minimized, with technical fixes simply being one possible, partial means to that end.

Relatedly, it is worth briefly addressing the offensive value of cybersecurity vulnerabilities, in particular relative to the ways in which algorithmic systems could be similarly exploited by malicious actors. The digitalization of global society since the 1990s has created innumerable opportunities for computer-enabled surveillance, disruption, expropriation, and illicit exchange of goods and information. As the 21st century has unfolded, both governments and organized, transnational criminal groups have shown a growing willingness to pay top dollar for tools, as well as associated tradecraft, that can enable the subversion of computer systems in accordance with their respective objectives. At least in part, the emergence of tools, processes, and mechanisms that are discussed throughout this paper for enabling defense-oriented exchange of vulnerability information have evolved in response to the emergence of offense-oriented markets for vulnerabilities and exploits. Much has been written on these topics, but readers should consider Perlroth (2021a) for a journalistic introduction and Ellis, Huang, Siegel, Moussouris, and Houghton (2017, 129-160) for a more focused overview of how BBPs can be constructed to disrupt the offensive side of the vulnerability market. To date, these dynamics don't

13 AJL Interview: Rayna Stamboliyska

translate well into the algorithmic harms space, where no similar offensive market for algorithmic vulnerabilities (wherein known flaws in algorithmic systems would be offered to the highest bidder for them to exploit) currently exists.[14]

"BUG BOUNTIES AND HACKERONE'S OWN MARKETING IS GUILTY OF … [DEALING] PREDOMINANTLY WITH TECHNICAL SECURITY VULNERABILITIES AND SECURITY DEFINED THROUGH A PRETTY NARROW SENSE. NOT BECAUSE THAT'S THE EXTENT OF WHERE [THE COMMUNITY] CAN PROVIDE VALUE, BUT IT COMES TO A QUESTION OF WHO IS INSTITUTING THESE PROGRAMS, AND ARE THEY IN A POSITION TO DO SOMETHING WITH THE FEEDBACK?"

— Alex Rice, CTO, HackerOne[15]

Overall, while important differences do exist between cybersecurity vulnerabilities and harmful flaws in algorithmic systems, ultimately both algorithmic harms and cybersecurity vulnerabilities represent serious problems to be documented, acknowledged, and addressed. The necessity of well-designed reporting and disclosure mechanisms for each is clear, regardless of precisely how the problems to be disclosed are defined.

TACKLING CYBERSECURITY VULNERABILITIES AND ALGORITHMIC HARMS

Both algorithmic harms and cybersecurity vulnerabilities can arise during the design, development, and/or deployment of sociotechnical systems. While the importance of considering this full arc appears, at least on paper, to be common wisdom in the cybersecurity space, we have found that mainstream discourse on algorithmic harms can at times ignore this fact and instead over-focus on specific parts of the product lifecycle, such as the collection of training data.

A Lifecycle Approach

As a result, we suggest that an approach similar to the cybersecurity best practice of integrating a "secure development lifecycle" (SDL) into overall product development may be usefully exported to the algorithmic harms context. Furthermore, "a critical step for continuous improvement of the SDL" (Rice et al. 2018, 32), which as a concept dates back almost two decades (Microsoft 2005), is the creation and identification of feedback loops — where the root causes of identified vulnerabilities are explored to iteratively update secure development practices. A similar approach for algorithmic systems — a 'Risk-based, Equitable, Accountable Lifecycle,' or 'REAL,' AI system design framework, as proposed by AJL's

14 A growing body of literature seeks to address what exactly these offensive uses of algorithmic systems could look like, with researchers at times highlighting that such concerns currently seem overblown. For instance, Tramèr (2021) reviews prominent categories of hypothesized algorithmic exploitation — algorithmic evasion, data inference, model poisoning, and model stealing — on the basis of their present, real-world impact, and concludes the following: "The strong attacks that we have developed in … research aren't really realistic in the vast majority of practical settings. There remains a big open question of what real attacks and defenses on real algorithmic systems-learning systems should look like" (Tramèr 2021). However, Brundage et al. identify various other ways in which algorithmic systems may be intentionally abused by malicious actors, similar to the ways in which cybersecurity vulnerabilities are exploited. For example, algorithmic systems may be used to automate social engineering attacks, manipulate information environments, and enable automated vulnerability discovery and exploit development (Brundage et al. 2018).

15 AJL Interview: Alex Rice

Director of Research & Design, Dr. Sasha Costanza-Chock[16] — could help formalize the many different points at which harms can be produced by an algorithmic system to drive consistent management of algorithmic system development and iterative, feedback-based improvement. Adopting such an overarching, lifecycle-based perspective may also help algorithmic system vendors and operators, as well as the field more broadly, move beyond the often myopic focus on biased training data or poor model choices. Similarly, consideration of cybersecurity incident or data breach response and reporting practices may present lessons for the algorithmic harms space, especially since such occurrences center on an experienced historic or ongoing harm, rather than merely the risk of harm implicated by vulnerability discovery.

"[THE] QUESTION AT THE HEART OF IT … IS, ONCE I'VE DISCOVERED A VULNERABILITY IN A PIECE OF SOFTWARE THAT I USE, WHAT SHOULD I DO WITH IT? WHO SHOULD I TELL ABOUT IT, IF ANYBODY, AND WHAT IS THE RIGHT SET OF ETHICAL, MORAL, AND COMMERCIAL CONCERNS WITH WHAT TO DO WITH THAT INFORMATION? BECAUSE THAT INFORMATION HAS IMPACT AND HAS VALUE AND HAS A WHOLE SUITE OF CONSIDERATIONS THAT ARE HARD FOR YOU AS AN INDIVIDUAL TO ASSESS."

— Alex Rice, CTO, HackerOne[17]

Underlying Business Processes and Incentives

Research focused on underlying causes for the occurence of both algorithmic harms and cybersecurity vulnerabilities highlights interesting parallels, especially regarding misaligned economic incentives faced by vendors operating in both spaces (Raji, Smart, et al. 2020; Hahn and Layne-Farrar 2006). Business incentives or plain convenience can outweigh individual or organizational intention to avoid harmful algorithmic impacts, just as these same forces often outweigh a desire to prevent or address cybersecurity vulnerabilities. Such imbalances manifest in a lack of accountability throughout the development lifecycle, including in approaches that allow new or updated products to be rushed to market without adequate scrutiny for cybersecurity or algorithmic harms. In particular, recent research from leading AI ethics scholars notes that algorithmic development "does not typically follow the waterfall or verification-and-validation approach" (Raji, Smart, et al. 2020, 37). Instead, an Agile development approach is typical, but while this is "much faster and iterative… [it] presents a challenge to auditability" (Ibid., 37). Similar dynamics, where product development is focused on speed at the expense of security, have long plagued the technology industry. Development and deployment processes that lack accountability and optimize for speed will continue to serve as breeding grounds for both cybersecurity vulnerabilities and algorithmic harms, in turn requiring robust mechanisms for holding vendors and operators to account retrospectively for such deficiencies.

16 Personal communication to authors from Sasha Costanza-Chock (2020)

17 AJL Interview: Alex Rice

"[A] PROBLEM THAT WE HAVE IS THAT INFORMATION SECURITY TOOLS, [SUCH AS] CVSS, … [ARE NOT DESIGNED FOR] FRONTIER CASES [SO] WE ARE TRYING … TO SHOEHORN NEW THINGS … INTO FRAMEWORKS THAT ARE VERY STRICT, THAT WE ARE VERY … COMFORTABLE WITH BECAUSE WE'VE BEEN USING THEM FOR AGES."

— Rayna Stamboliyska, Fmr. VP Governance and Public Affairs, YesWeHack[18]

Scale and Complexity

Scale and system complexity also present challenges in both cases. As Raji and Smart et al. note, a dynamic of growing complexity compounding the challenge of adequate oversight has historic parallels within the aerospace and financial services industries, which "had to play catch-up as the complexity and automation of … business practices became too unwieldy to manage manually" (2020, 36). In the cybersecurity context, according to Zhao, Laszka, and Grossklags, "[f]eature-rich complex software products are inherently difficult to develop securely, in particular, if time-to-market … and other business realities impede careful engineering practices" (2017, 373).

Flawed, Legacy Inputs

Legacy design inputs in both security and algorithmic development contribute directly to these issues. For algorithms, the use of outdated datasets, rife with problematic biases and noisy labels, parallels the influence of obsolete legacy functions and languages in modern software, which were developed without sufficient attention to security. One example of an enduring source of harm in algorithmic systems is CoNLL-2003, which is "relied on as an evaluation tool to validate some of the most-used language systems" (Field, 2020). A model trained on this nearly twenty-year old dataset "would have more trouble with women's names, but it would also likely be worse at recognizing names more common to minorities [sic], immigrants, young people, and any other group that wasn't regularly covered in the news two decades ago" (Ibid.). Despite the deficiencies in CoNLL-2003, its open-source availability and widespread use continues to make it a "yardstick" for the validation of language recognition algorithms (Ibid.). Meanwhile, in the cybersecurity domain, various programming languages still in widespread use today (e.g., C, C++, JavaScript, PHP) "have functions … whose security implications were not appreciated when initially introduced but are now widely regarded as dangerous" (Rice et al. 2018, 15). At the same time, business-driven requirements to maintain interoperability of new and legacy systems are often met in ways that sustain vulnerabilities across product versions, since "once … developers understand that the older, less secure technology is allowed to live on, solutions that would ease the risk are often not considered at all" (Zatko 2009).

"[IF] YOU'RE LOOKING FOR THESE KINDS OF [ALGORITHMIC] HARMS, [THEN] IT COULD IMPLICATE SIMILAR LAWS. FOR INSTANCE, DMCA, WHENEVER YOU'RE REVERSE ENGINEERING SOFTWARE."

— Jack Cable, Security Researcher[19]

18 AJL Interview: Rayna Stamboliyska

19 AJL Interview: Jack Cable

Third-Party Audits

Both algorithmic harms and cybersecurity vulnerabilities can be identified and analyzed through audit-like mechanisms. Although an imperfect and incomplete solution, these are both a well-established practice in cybersecurity and increasingly popular for addressing algorithmic harms (Johnson 2021). These kinds of external audits, when conducted by credible third-party researchers — either independently or working on behalf of an auditing organization — are "less affected by organization-internal considerations" that could otherwise hamper end-goals of transparency and accountability (Raji, Smart, et al. 2020, 35). To the extent that these approaches involve "seeking inputs from diverse populations and forcing [internal teams] to try to refute [their] assumptions," they can also be beneficial in helping organizations to escape so-called "confirmation traps" (Zatko 2009). Absent such diversity, testing approaches, as a result of deep-rooted human biases, tend to be designed to validate preconceptions of how the systems' designers intended them to be — 'the network is secure' or 'the model is unbiased' — rather than to scrutinize those systems and assumptions through a truly adversarial lens (Ibid.). At present, however, independent, third-party testing for algorithmic harms is "limited by lack of access … at the audited organizations" to underlying technology or details of relevant business processes (Raji, Smart, et al. 2020, 35), since end-user-level interactions with algorithmic systems alone are rarely sufficient for comprehensive testing and validation of those systems' oft-presumed adequacy (Sandvig et al. 2014).

"I STARTED TO WORK ON THE ALGORITHMIC BUG BOUNTY CONCEPT MOSTLY BECAUSE OF THE SHARED … ISSUES THAT SECURITY RESEARCHERS MIGHT EXPERIENCE, AND THE POTENTIAL ISSUES THAT ALGORITHMIC AUDITORS MIGHT EXPERIENCE. THERE ARE SOME AREAS WHERE THERE ARE SIMILAR … POTENTIAL ISSUES [FOR EXAMPLE] AROUND THE CFAA AND THE ANTI-HACKING LANDSCAPE [AND] THERE ARE SOME AREAS WHERE THEY'RE DIFFERENT."

— Amit Elazari Bar On, Director, Global Cybersecurity Policy, Intel Corporation[20]

Protecting Independent Researchers

Beyond practical impediments that exist to independent researchers being able to freely access all relevant details of systems to be audited, undertaking such examinations may expose researchers to the risk of legal retribution by vendors or operators. As alluded to in the Introduction, institutions that seek to avoid scrutiny for their harmful algorithmic systems or lax security practices may seek to use the law to suppress the voices of those who would hold them to account. Fortunately, in practice, legal conflicts in the vulnerability research space rarely escalate beyond the "saber rattling" issuance of cease-and-desist letters from the relevant vendor or operator.[21] However, the global patchwork of national, state, and local laws can still create a hard-to-navigate landscape for researchers, wherein the line between legal and illegal research activities is inconsistent across contexts and jurisdictions. It is the breadth and

20 AJL Interview: Amit Elazari Bar On

21 AJL Interview: Marcia Hofmann

complexity of legal risks, therefore, rather than any one provision or statute, which pose the greatest challenge to researchers. For example, even looking solely at federal U.S. law, researchers must navigate the intricacies of standard copyright law, contract law, trade secret law, privacy law, and export control law (Park and Albert 2020). Defamation and extortion can also come to play when disclosures are involved (Whittaker 2018). Given that the threat of legal action acts a deterrent to both the process and publication of security research (Gamero-Garrido et al. 2017), the question of how to manage and mitigate legal risk, both real and perceived, and thereby avoid 'chilling effects' on research activities, has been unsurprisingly front-and-center in the development of the vulnerability research ecosystem over the past two decades. The widespread adoption over recent years of 'legal safe harbor' clauses in BBP program terms as a direct response to this concern has afforded a greater degree of protection for researchers from civil legal action (and, for government BBPs, criminal action).[22] However, according to Amit Elazari Bar On, one of the foremost experts on the legal aspects of BBPs, "algorithmic auditors and security researchers have much in common when it comes to dealing with US anti-hacking law's murky landscape" (2018), hence, it is reasonable to presume that similar consideration and innovation in legal protections for researchers will be needed in the algorithmic harms space.

"ONE OF THE THINGS THAT I IMAGINE WOULD BE TRICKY IS THAT [ALGORITHMS] ARE TREATED AS PROPRIETARY, SO TRYING TO FIGURE OUT HOW TO [EMULATE THE APPROACHES FROM CYBERSECURITY] IN A WAY THAT WOULD MAKE THESE COMPANIES COMFORTABLE ABOUT SUBMITTING PROPRIETARY INFORMATION FOR SCRUTINY WOULD BE AN ISSUE."

— Marcia Hofmann, Digital Rights Lawyer[23]

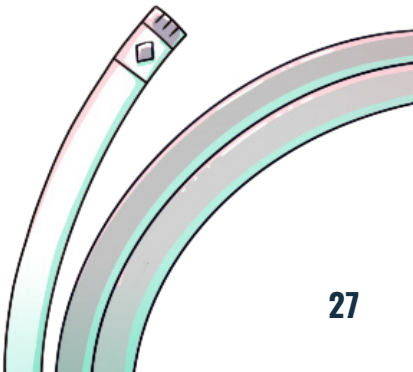
Reporting and Assessment

Lastly, both in cybersecurity and for algorithmic harms, it is important to note that risks of harm vary widely with the context of system use, and questions around measurement and documentation of causal mechanisms and impacts are therefore paramount. Algorithmic systems can induce or entrench existing inequities when they are used to determine individual and aggregate outcomes in areas including criminal justice, healthcare, education, and employment (Altman, Wood, and Vayena 2018). In the cybersecurity space, the anticipated impact of vulnerabilities is dependent on context: where is the vulnerable system deployed; who may want to exploit the vulnerability; which users and organizations are likely to be impacted; and so on. To a certain extent, this context is captured through vulnerability assessment frameworks, such as the Common Vulnerability Scoring System (CVSS), which provide otherwise standardized approaches for severity measurement across different vulnerabilities (FIRST 2019). Once scored, each newly discovered vulnerability is assigned a unique identifier ("Common Vulnerability Enumeration," or "CVE") that allows for consistent

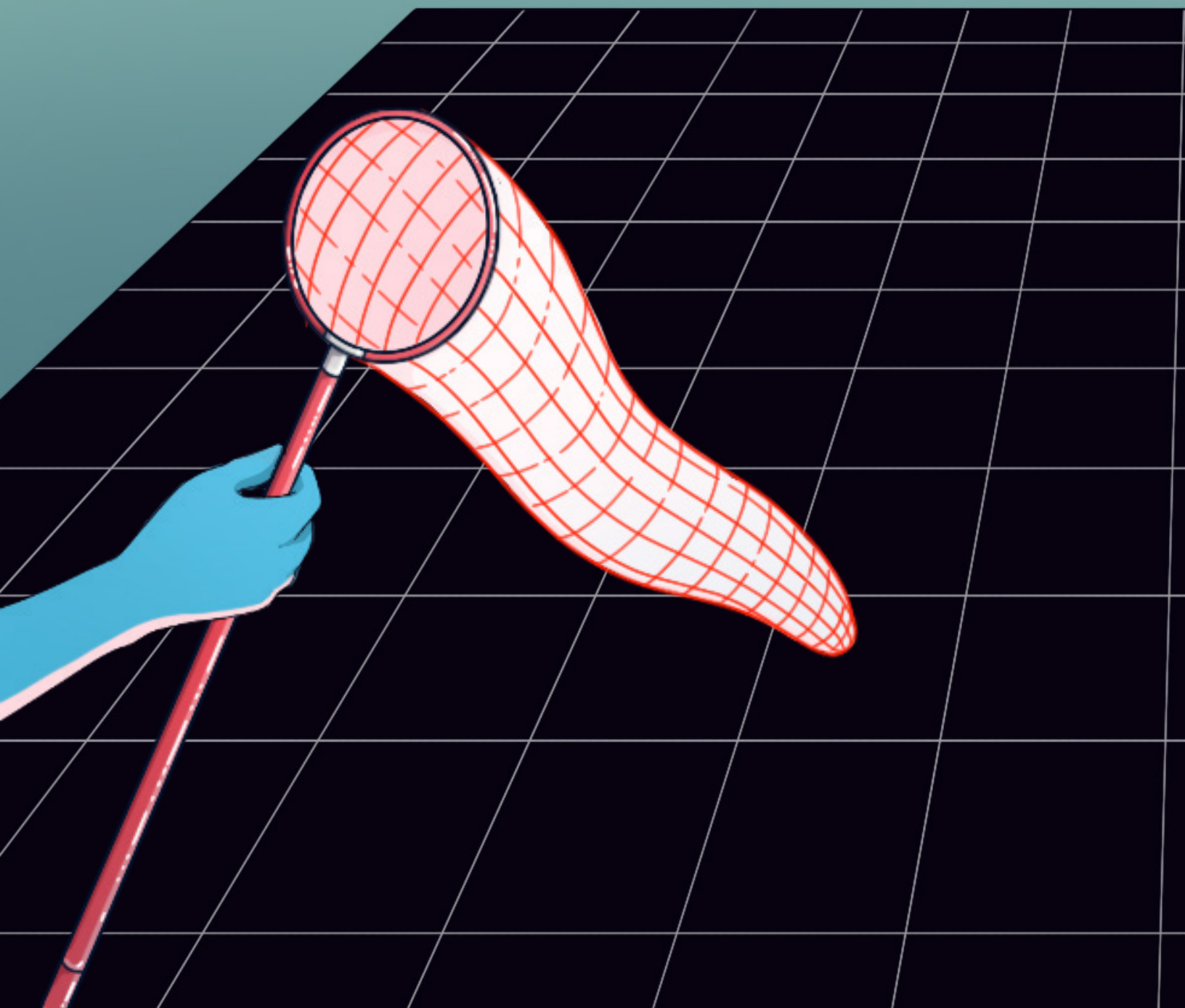
22 AJL Interview: Amit Elazari Bar On

23 AJL Interview: Marcia Hofmann

tracking and identification of that vulnerability across databases (e.g., U.S. “National Vulnerability Database,” or “NVD”), systems, and organizations. Meanwhile, other tools — such as ExploitDB and VirusTotal — exist to track exploits built upon particular vulnerabilities. A similar assortment of mechanisms to capture, score, and track algorithmic harms, as well as their causes, could well be useful as efforts to address those issues become more consistent and systematized across organizations (refer to Design Lessons from BBPs for more on this topic).



III.
BUG BOUNTY PROGRAMS 101



THE EMERGENCE OF 'THE' BUG BOUNTY MODEL

"TO ME, THE ORIGIN OF THIS COMMUNITY AND THIS PRACTICE BEGINS WITH VULNERABILITY DISCLOSURE, WHICH HAS EXISTED IN SOME FORM OR ANOTHER, AN EVOLVING FORM FOR AS LONG AS THE INTERNET HAS, AS LONG AS PEOPLE HAVE BEEN PRODUCING SOFTWARE."

— Alex Rice, CTO, HackerOne[24]

The earliest BBPs for cybersecurity vulnerabilities — such as Netscape's 1995 "Bugs Bounty" (Ellis and Stevens Forthcoming), and iDefense's 2002 "Vulnerability Contributor Program" (Zorz 2003) — were established at a time in history when relations between independent security researchers and technology vendors and operators were at a low ebb. Undoubtedly, one reason for this situation was the underlying disposition of vendors and operators towards security during that era, in that security assurance and oversight was largely an afterthought in the development of new products and product functionality (Anderson and Moore 2006). Meanwhile, hacking and hacking-adjacent communities had been proliferating and, with a wide variety of motivations, exploring and exposing security deficiencies. However, rather than perceiving these groups for the motivationally diverse communities which they were, or seeing the potential value of collaborating with them to help address widespread insecurity, key decision-makers in the private sector and in government largely perceived hackers as a monolith of misanthropy, cultural deviance, and criminality (Nissenbaum 2004, Goerzen and Coleman Forthcoming). At least as far back as the 1980s, hacking communities had been portrayed in such a manner by mainstream institutions from Hollywood to Capitol Hill, and journalism to law enforcement (Ibid.), with this flattened misrepresentation in-part motivating the passage of anti-hacking provisions within the DMCA (Electronic Frontier Foundation 2014) and the CFAA (Schulte 2008), as well as the CFAA's subsequent scope creep (Paolillo 2017). Ironically, even well-intentioned, sympathetic portraits of hackers and their supposedly universal 'ethic' — summarized by Gabriella Coleman as "a commitment to information freedom, a mistrust of authority, a heightened dedication to meritocracy, and the firm belief that computers can be the basis for beauty and a better world" (2012, 99) — share this through-line of overgeneralization. This portrait of a universal "hacker ethic" obscures the fact that, "similar to any cultural sphere, we can easily identify variance, ambiguity, and, at times, even serious points of contention" across hacking communities (Ibid.).

"[YOU'RE] USING STANDARD FORM CONTRACTS ON A LARGE SCALE TO CREATE AND PROMOTE ADOPTION OF A SET OF PROVISIONS THAT MIGHT NOT ADDRESS ALL THE [POTENTIAL LEGAL] ISSUES, BUT ARE ADVANCING … THE LEVELS OF CERTAINTY IN A WAY [THAT IS] VERY SIMILAR TO … CREATIVE COMMONS [FOR COPYRIGHT CLAIMS]."

— Amit Elazari Bar On, Director, Global Cybersecurity Policy, Intel Corporation[25]

24 AJL Interview: Alex Rice

25 AJL Interview: Amit Elazari Bar On

"WE WERE IN A POSITION WHERE THE WELL WAS POISONED. [SOME OF THE] RESEARCHER COMMUNITY … HAD BEEN ISSUED CEASE AND DESIST NOTICES [FOR] … SIMPLY FOR TRYING TO SUBMIT INFORMATION TO [THE DOD] BEFORE. AND I DON'T KNOW ABOUT YOU, BUT IF EVER I WERE TO RECEIVE SOMETHING LIKE THAT, IT [WOULDN'T] MATTER HOW MANY OLIVE BRANCHES HAD BEEN EXTENDED FOR ME, I PROBABLY [WOULD] NEVER GET OVER THAT AND TRY TO HELP AGAIN … [THE] HACKER COMMUNITY, WHETHER IT'S [THROUGH] POP CULTURE OR THE GOVERNMENT OR WHOEVER ELSE, … THEY WERE REALLY THOUGHT OF AS CRIMINAL."

— Lisa Wiswell Coe, Fmr. Program Manager, Hack the Pentagon[26]

By the late 1990s, the hostile legal backdrop faced by defense-oriented hackers and other independent security researchers[27], coincident with vendors and operators generally following a PR-first "security through obscurity" approach, helped to produce a situation in which vulnerability disclosures were frequently uncoordinated, public, and messy, but also the only feasible means for these individuals to compel action by vendors and operators to remedy those flaws (Schneier 2007). Researchers incurred consequential legal risk under the murky auspices of assorted anti-hacking laws, even as their work in researching and drawing attention to individual security weaknesses and flawed design and business practices more generally went uncompensated and, often, under-appreciated.[28] All told, and as discussed throughout this paper, that historic state of affairs encompassed many thorny and difficult dynamics similar to those present in the contemporary algorithmic harms research ecosystem.

"[A]CCOUNTABILITY WAS A THEME, EVEN GOING BACK TO THE LATE 90'S. A LOT OF PEOPLE ENGAGING IN [SECURITY RESEARCH] THOUGHT THAT SOFTWARE WASN'T GOOD ENOUGH, AND THEY WANTED ACCOUNTABILITY FROM THE MANUFACTURERS [AND] CREATORS OF THAT SOFTWARE [THAT] YOU [AS A USER] ARE POWERLESS TO IMPROVE. THE ONLY OPTION YOU HAVE TO PROTECT YOURSELF IS THE ACCOUNTABILITY OF THAT COMPANY."

— Dino Dai Zovi, Security Researcher[29]

Today, the overall state of the vulnerability disclosure landscape is vastly improved from that historic nadir, a fact that in itself partially served to motivate AJL's initial interest in this area and the potential of models from the cybersecurity domain to inform the creation of mechanisms for surfacing and redress of algorithmic harms. Community-led efforts have featured prominently in this transformation. The

26 AJL Interview: Lisa Wiswell Coe

27 As discussed above in Tackling Cybersecurity Vulnerabilities and Algorithmic Harms, and throughout this paper.

28 Notwithstanding the derision that hackers of this era received for high-profile public disclosures of vulnerabilities, Goerzen and Coleman (Forthcoming) argue that this "security by spectacle" approach had important implications for the establishment of the field of cybersecurity: "[The] hacker-led process of disclosure and attention-seeking functions as a case study for a novel mode of what critical security scholars term "securitization": the process whereby powerful institutional actors, like the state and massive commercial entities, deem a particular issue to be an extraordinary threat, and thus to warrant extraordinary measures of address through security processes."

29 AJL Interview: Dino Dai Zovi

earliest efforts to document consistent pathways for vulnerability disclosure, for example by Jeff Forristal (aka. "Rain Forest Puppy" in hacker circles), laid the conceptual groundwork for future VDPs and BBPs (Goerzen and Coleman Forthcoming). In addition, the titular objective of the researcher-led "No More Free Bugs" initiative, unveiled to the world at the CanSecWest security conference in 2009 (Dai Zovi, 2009), has been largely accomplished in the eyes of its co-founder, security researcher Dino Dai Zovi, although with a focus on vulnerabilities in web-facing services, rather than traditional hardware or software, that he had not anticipated at the time.[30]

Meanwhile, several factors have contributed to an overall lowering of legal risk to security researchers, including improved communication of research expectations by vendors and operators; robust advocacy in defense of security researchers and their interests; the adoption of legal safe harbor clauses in BBP rules;[31] the articulation of security research principles and best practices by a number of governments (Stevens et al. 2021, 41-42); and constraints on the scope of DMCA and CFAA provisions impacting security research, by the U.S. Library of Congress (Geiger 2018) and Supreme Court (Mackey and Opsahl 2021), respectively. At the same time, albeit often motivated by a combination of BBPs' security and PR value, a wide variety of technology vendors and operators are now inviting independent security researchers to scrutinize their systems in order to meet a rising tide of customer and government expectations around security. However, despite this progress, an aversion on the part of vendors and operators to truly adversarial research and meaningful transparency through the consistent public disclosure of vulnerabilities has not only been allowed to endure, but arguably has been further entrenched via the mechanisms through which independent security research has been institutionalized.[32]

CONTEXTUALIZING BBPs AS VULNERABILITY REPORTING MECHANISMS

"YOU ARE LOOKING FOR A DEFINED PROBLEM THAT CAN HAVE A FIX. IF IT DOESN'T HAVE A FIX, THEN WHY FIND IT? … WE RECEIVE THOUSANDS OF SUBMISSIONS EVERY WEEK AND WE VALIDATE THEM. AND IF THEY'RE POSSIBLE TO VALIDATE BECAUSE THEY'RE REPRODUCIBLE, THEY'RE STILL THERE UNTIL FIXED."

— Mårten Mickos, CEO, HackerOne[33]

Looking beyond the prototypical BBP, we have observed a great deal of programmatic variety in how mechanisms for third-party research and disclosure of cybersecurity vulnerabilities have been implemented to date, driven by competitive pressures of growing demand for both security

30 AJL Interview: Dino Dai Zovi

31 AJL Interview: Amit Elazari Bar On

32 Refer to Key Takeaway #5 for further context.

33 AJL Interview: Mårten Mickos

research talent and vulnerability information from a wide range of organizations.[34] In order to better understand the full breadth of this ecosystem, and to illustrate where BBPs diverge from other vulnerability reporting approaches, we have developed a typology of design levers that can be used to distinguish these various mechanisms (Figure 1).[35] In doing so, we focused on a number of key differences across programs, namely:

- **Target Entities** — Does a particular program or platform solicit reports only for vulnerabilities affecting organizations that have agreed to receive such reports, or can reports be submitted even for organizations that do not agree to participate?
- **Compensation Model** — How are security researchers compensated for their contributions and expertise?
- **Disclosure Model** — Under what terms are researchers who discover a vulnerability authorized to publicly disclose their findings (e.g., to the press, in blog posts, or academic research), and on what timeline?
- **Participation Model** — To what extent is the program intended for widespread, public participation ("open") versus permitting only a numerically limited number of select researchers ("closed")?
- **Program Management** — To what extent are the responsibilities of program management handled directly by target organizations versus by third-party platforms? (e.g., hosting program terms, receiving reports, validating submissions, triaging vulnerabilities, facilitating patch development, and verifying patches).
- **Program Duration** — Are programs intended to be temporary, or long-lived?
- **Program Scope & Access** — How wide is the range of systems and vulnerability types formally encompassed under a given program, and what level of technical and organizational access is afforded to researchers to facilitate exploration and analysis?

"THERE'S NOTHING IN THE LAW THAT SAYS THAT A BUG BOUNTY PROGRAM IS A LEGAL SAFE HARBOR. LIKE THERE'S NO CARVE OUT IN THE COMPUTER FRAUD AND ABUSE ACT. THERE'S NOTHING IN, FOR EXAMPLE, COPYRIGHT LAW THAT WOULD SAY … IF YOU'RE ACTING PURSUANT TO THE TERMS OF THE BUG BOUNTY PROGRAM, THERE'S A LEGAL SAFE HARBOR, THERE'S NOTHING LIKE THAT."

— Marcia Hofmann, Digital Rights Lawyer[36]

Different combinations of choices for these design levers can be used to generate various vulnerability reporting mechanisms and platforms that differentially motivate and target security

34 Definitions for BBPs and vulnerabilities can be found in the Objectives sub-section of the Introduction and Comparability of Vulnerabilities & Algorithmic Harms, respectively.

35 In part this framework builds upon variations in "program configuration" noted by Ellis, Huang, Siegel, Moussouris, and Houghton: "Programs vary in several ways, including how they define market access (who can participate as a seller), program duration (when will sales be accepted), and compensation (what is offered as a reward)" (2017, 133).

36 AJL Interview: Marcia Hofmann

researchers. In turn, these mechanisms can be used to advance distinct organizational priorities. For example, an organization might operate a private, time-bounded BBP that is narrowly focused on a beta release of a new product, requires researchers to adhere to non-disclosure agreements, affords the selected researchers access to product source code, and offers consistently high rewards across vulnerability types. Such a program would be effective for directing the attention of vetted researchers with specific, established skills towards improving the security of this particular product ahead of its release to the public. Concurrently, that same organization may also offer a public, ongoing BBP that covers a wide range of products and internet infrastructure, allows for public disclosure only by mutual agreement, and offers a range of monetary and non-monetary rewards across different vulnerability types, but doesn't afford any privileged access to underlying code, in order to achieve more marginal, diffuse, and, from a technical perspective, more surface-level security improvements. Importantly, we have found these levers to be sufficiently generalizable that they would translate to programs focused on broader socio-technical harms, such as algorithmic harms.

This framework enables us to qualitatively compare vulnerability disclosure mechanisms, including public and private BBPs, and to illustrate commonalities and differences, as shown in Figure 2. In general, more 'open' mechanisms (i.e. those that allow for public participation, rather than pre-vetting a limited number of contributors) offer a lower level and reliability of compensation, and tend not to afford additional technical or organizational access, but also allow researchers to retain greater discretion over how reported information may be distributed (i.e. full or coordinated versus limited or non-disclosure). These mechanisms can be construed as closer to a 'crowdsourcing' model, and also tend to welcome reports on a wider range of vulnerabilities or infrastructure, although limitations on access may in practice constrain the variety of vulnerabilities submitted and systems covered. In contrast, 'closed' mechanisms that are tailored towards more selective participation tend to limit public disclosure more extensively, provide more reliable compensation (e.g., through a contract stipulating up-front payment), and provide deeper access while focusing on specific elements of organizations' infrastructure or particular types of security weakness. Because of these various points of design divergence, we do not consider BBPs as a well-bounded category but rather as one design configuration on a spectrum of vulnerability disclosure mechanisms, the core characteristics of which can be abstracted to identify design lessons applicable to algorithmic harms. Vulnerability disclosure programs ("VDPs") typify the 'open' configuration described above, while penetration testing ("pentesting") typifies the 'closed' configuration. Public and private BBPs typically sit in between.

"I THINK … THE REASON SAFE HARBOR … HAS GARNERED RESULTS [IS BECAUSE] IT WAS ALSO LOOKING FOR A SOLUTION … FOCUSED ON PRIVATE ORDERING. IT WAS ACHIEVABLE BECAUSE IT WAS FOCUSED ON FLEXIBLE CONTRACTUAL LANGUAGE [THAT] COULD BE ADAPTED FAIRLY EASILY AND RAPIDLY, AS OPPOSED TO OTHER TYPES OF POLICY REFORM, WHICH ARE OF COURSE STILL VERY VALUABLE."

— Amit Elazari Bar On, Director, Global Cybersecurity Policy, Intel Corporation[37]

37 AJL Interview: Amit Elazari Bar On

Organizations may adopt one or more of these mechanisms, as their design distinctions make them better suited to different moments in product or infrastructure development, differing levels of organizational security maturity, or meeting distinct cybersecurity goals. For example, abstracting from our research into BBPs, 'closed' participation models tend to yield a lower overall volume of vulnerability reports but also decrease "noise" (duplicative, inaccurate, or nonsense reports), making them a better fit for organizations just beginning to build cybersecurity capacity and process maturity internally (HackerOne 2019a, 5). In contrast, for organizations whose cybersecurity practices are already robust, those same mechanisms may be useful for testing sensitive systems (e.g., non-public infrastructure) or products prior to public release.

"THERE IS … A DISTINCTION TO BE MADE … BETWEEN A COMPANY THAT HAS A COORDINATED VULNERABILITY DISCLOSURE POLICY, WHERE THEY'RE NOT PAYING ANYTHING, [AND] WHERE THEN IT'S A LOT CLEARER THAT YOU SHOULD BE ABLE TO DISCLOSE THE VULNERABILITY AFTER A SET NUMBER OF DAYS, OR ONCE IT'S BEEN FIXED, VERSUS WHEN A COMPANY HAS A BUG BOUNTY PROGRAM [AND] THEY'RE PAYING YOU … THEN IT'S A LITTLE MUDDIER BECAUSE YOU'RE GETTING SOME BENEFIT OUT OF THAT."

— Jack Cable, Security Researcher[38]

38 AJL Interview: Jack Cable

Figure 1: Design Levers Across the BBP Spectrum

TARGET ENTITIES	**Voluntary**		**Adversarial**	
	Only reports relating to organizations that have consented to receiving such reports are accepted.		Reports related to organizations that have not agreed to receive such reports are also accepted.	
COMPENSATION MODEL	**Non-Monetary**	**Bounties**	**Contract**	**Employment**
	Security researchers receive only non-monetary benefits in exchange for their findings.	Organizations or platforms pay security researchers for finding in-scope vulnerabilities, with rediscovery of already-identified vulnerabilities genverally not rewarded.	Organizations or platforms retain security researchers on a temporary, contractual basis to undertake specific services. Researchers are compensated regardless of findings.	Organizations or platforms retain security researchers on a permanent basis to undertake wide-ranging research. Researchers are salaried and receive typical employment benefits.
DISCLOSURE MODEL	**Delayed Full Disclosure**	**Coordinated Disclosure**	**Non-Disclosure**	
	Researchers can freely disclose their findings to the public on a predetermined time frame without additional approval from the affected organization.	Organizations contract with a BBP platform to provide specific services within or related to their BBP, while handling other elements in-house.	Researchers cannot publicly disclose their findings.	
PARTICIPATION MODEL	**Public**		**Private / Invite-Only**	
	All researchers are invited to conduct research and submit reports.		Only pre-authorized researchers are invited to conduct research and submit reports.	
PROGRAM MANAGEMENT	**Platform-Managed**	**Mixed Management**	**Self-Managed**	
	Organizations contract with a BBP platform to deliver their BBP, with reports typically channeled through a specific portal on the platform. The platform also provides related services (e.g., report validation, triage, patch verification, etc.).	Organizations contract with a BBP platform to provide specific services within or related to their BBP, while handling other elements in-house.	Organizations handle delivery of their BBP in-house. They accept reports directly through their websites or via a dedicated email address and handle validation, triage, andpatch verification internally .	
PROGRAM DURATION	**Ongoing** ⟷		**Time-Limited**	
	Reports are accepted on a continuous basis for an evolving range of assets.		Reports are accepted for a specified range of assets over a limited time period.	
PROGRAM SCOPE & ACCESS	**Constrained** ⟷		**Expansive**	
	Only a limited variety of vulnerability types and / or systems are identified as in-scope of the program.		All possible vulnerability types and systems are in-scope of the program.	
	‘Closed Box’		**‘Open Box’**	
	Testing is limited to publicly available resources and tooling, without additional organizational or technical access (e.g., to documentation or source code).		Additional access, either organizational or technical, is provided to enable a deeper level of testing.	

Figure 2: Taxonomy of Typical Vulnerability Reporting Mechanisms

	VULNERABILITY DISCLOSURE PROGRAMS (VDPs)	PUBLIC BBPs	PRIVATE BBPs	PENETRATION TESTING
COMPENSATION MODEL	Non-Monetary	Bounties	Bounties	Contract / Employment
TYPICAL DISCLOSURE MODEL	Delayed Full Disclosure or Coordinated Disclosure	Coordinated Disclosure or Non-Disclosure	Non-Disclosure	Non-Disclosure
PARTICIPATION MODEL	Public	Public	Private / Invite-Only	Private / Invite-Only
PROGRAM DURATION	Ongoing	Ongoing	Ongoing / Time-Limited	Time-Limited
EXAMPLES (PRIVATE SECTOR)[39]	Capital One	TikTok	LinkedIn	Rapid7 Penetration Testing Services
EXAMPLES (U.S. GOVERNMENT)	VDP Template for U.S. Government Agencies[40]	General Services Administration TTS Bug Bounty	Hack the Army 3.0	Cloud.gov Penetration Test Authorization

A Note on Program Scope and Access

Pentesters and vetted researchers in private BBPs tend to be afforded greater access than unvetted participants in public BBPs or contributors to VDPs. An explanation for this lies with the concept of 'trust.' Exposing the details of high-value systems or systems with regulatory or reputational significance can create risks for an organization (e.g., from misappropriation or misuse), while building and maintaining trust is challenging as participation scales. Hence, an affordance of greater access tends to be limited to these more selective programs. However, the interaction of this dynamic with the financing aspects discussed in Key Takeaway #5, wherein target organizations hold the reins of the ecosystem and thus ultimately control both program scope and disclosure terms, leaves much to be desired. For a comparable mechanism to facilitate compensated, adversarial research into algorithmic systems, including 'open box' testing, alternative approaches to vetting participants and enabling access are needed.

A Note on Full Disclosure

The concept of "full disclosure" encompasses a variety of approaches and motivations, and has been the topic of enduring debate across security-focused communities and organizations for decades. Goerzen and Coleman (Forthcoming) offer a comprehensive treatment of the term's complex history. The concept has seen proponents and detractors, arguing over trade-offs that are summarized by Matt Goerzen as follows:

- Accountability;
- Advertise vulnerability and facilitate exploitation in the absence of a patch;
- Provide knowledge of exploitation methods to malicious hackers, as well as other motivated actors; and
- Enable third-party security service providers to function as protection rackets.[41]

For the purposes of this paper, we chose to focus on an interpretation of full disclosure that straddles the high-level contours of debate on this topic; namely, a guarantee of comprehensive, public disclosure of information about discovered vulnerabilities on a predetermined time frame, captured as "delayed full disclosure." This approach allows adequate time for notification of the responsible vendor or operator, and that organization's development of a fix, assuming appropriate prioritization of attention and resources. Such an approach could be adopted in the algorithmic harms space, with variable treatment of timing considerations based upon circumstances, including the generally low risk posed by prompt publicization, relative to non-trivial security risks associated with the dissemination of cybersecurity vulnerability information before a fix has been developed.

39 The private sector VDP, private BBP, and public BBP listed as examples in Figure 2 are all managed and hosted by HackerOne. Although similar examples could have been drawn from other platforms (e.g., Bugcrowd, YesWeHack, etc.), we opt for a single source to allow readers to more easily compare the program terms for these different mechanisms.

40 In September 2020, the U.S. Cybersecurity and Infrastructure Security Agency (CISA), which is responsible for ensuring the cybersecurity of many U.S. government systems, released "Binding Operational Directive 20-01," requiring all executive branch agencies to establish VDP mechanisms for receiving vulnerability reports. The VDP template listed above was produced by CISA to facilitate implementation of this policy (Cybersecurity and Infrastructure Security Agency 2020). We believe that similar administrative approaches could be useful for addressing algorithmic harms.

41 Notes from conversation between the authors of this paper and Matt Goerzen (September 2021).

COMPARING REPORTING & DISCLOSURE PLATFORMS

When considering participatory reporting and disclosure mechanisms that can inspire methods to tackle algorithmic harms, we focused not only on individual programs (BBPs, VDPs, pentesting) but also specifically on the role that intermediary organizations — or platforms — could play in this regard. This section lays out initial findings on key design principles across a variety of intermediating platforms, which we obtained by applying the relevant elements of the typology described above.

"[ONE] REASON WHY I THINK IT'S CRITICAL TO DISTINGUISH BETWEEN A VDP AND A BBP IS BECAUSE … THEY'RE TRYING TO SOLVE VERY DIFFERENT ISSUES. … YOU CAN RUN, AT THE SAME TIME, A PUBLIC BUG BOUNTY WITH … ONE SCOPE IN MIND, AND A PRIVATE BUG BOUNTY … LOOKING, FOR EXAMPLE, [AT] BETA VERSIONS THAT YOU HAVEN'T EVEN RELEASED TO THE PUBLIC. … MY POSITION IS THAT YOU SHOULD HAVE BOTH."

— Amit Elazari Bar On, Director, Global Cybersecurity Policy, Intel Corporation[42]

Given our desire to be inclusive of the historic variety of these kinds of platforms, as well as the fact that modern platforms even ostensibly focused on BBPs still typically offer VDPs or pentesting services, we group all such platforms together in this section under the moniker of "reporting and disclosure platforms" (refer to Glossary for definition)..[43]

Looking at the archetypal platform for private BBPs, Synack, we observe that vulnerability reports for its private BBP (or "crowdsourced pentest") offerings are solicited only on behalf of organizations that sign up to receive such reports and are kept confidential by default (Synack, n.d., a). Security researchers that perform these services are part of the "Synack Red Team" and are paid under a bounty-type arrangement for their discoveries (Synack, n.d., b). Like many competitor platforms, Synack also maintains a general-purpose VDP for its researchers to report vulnerabilities affecting vendors outside of the Synack ecosystem (Synack 2015), as well as a range of other crowdsourced security testing options. HackerOne, while also hosting private BBPs and VDPs, emphasizes its position as the largest public reporting and disclosure platform by number of registered participants, and also pays security researchers under a bounty model (HackerOne 2021). Different public BBPs on the HackerOne platform have different disclosure rules, but only a small fraction of its public BBPs provide a clear and consistent timeframe under which researchers can disclose their findings.[44] Rather, the business models of prominent platforms offering BBPs depend on

42 AJL Interview: Amit Elazari Bar On

43 In analyzing these platforms, we suggest that a categorical distinction between "open" and "closed" platforms, defined in the same manner as in the section above, is worthy of particular emphasis. To facilitate that focus, Figure 3 places archetypal "open" and "closed" platforms — HackerOne and Synack, respectively — in context alongside other platforms delivering BBPs, VDPs, pentesting, and related mechanisms. Besides the two archetypal platforms, the other entities shown in Figure 3 were selected based on a desire to minimize the number of entities included while maximizing the programmatic distinctions between them. For example, while BugCrowd is a prominent reporting and disclosure platform, it offers largely similar services to clients, provides a comparable engagement experience for security researchers, and is funded through the same venture capital-backed approach as its competitor HackerOne. Hence, we can reasonably capture the offerings of both organizations through our inclusion of HackerOne as a "representative" platform that can reasonably be viewed as a stand-in for all organizations of that same model, at least at the level of abstraction that we employ within this framework. While this is a somewhat subjective means of sampling platforms, such an approach allows us to most easily distinguish between archetypal reporting and disclosure platforms and platforms or projects that deviate from those typical models.

44 Refer to Footnote 87 for details.

establishing and maintaining a growing number of productive relationships with client organizations. As such, the functions that these platforms play in intermediating disclosure and facilitating monetary compensation for researchers can only be applied to organizations that volunteer their systems for scrutiny; what we refer to in this paper as "non-adversarial" mechanisms.

In contrast, public, adversarial models that stipulate delayed full disclosure of all reported vulnerabilities (WooYun, Open Bug Bounty) have to date failed to provide the same level and consistency of compensation for researchers as public, voluntary BBP models. Anecdotal evidence also suggests that report quality and validity for Open Bug Bounty may be questionable (WordPress Support Forum 2020-2021). In effect, these mechanisms seem to end up as little more than VDPs with an aspirational but usually unmet goal of rewarding researchers for their findings. In contrast, Trend Micro's ZDI — a self-described "vendor-agnostic bug bounty program" (Trend Micro 2021) — is open to public participation but manages to provide generous rewards for vulnerability reports, as well as various other opportunities to garner financial bonuses. In part, the more generous nature of ZDI's rewards reflect its exclusive focus on high-value "zero-day" vulnerabilities, which as genuinely novel security flaws tend to be of significant value to nation-state hacking groups for offensive use, as well as — in the case of ZDI — vendors and operators looking to fix or mitigate or mitigate those vulnerabilities. However, like Google's Project Zero, which leverages in-house research talent rather than independent researchers, the ZDI model ultimately only works because of sponsorship by a corporate benefactor — Trend Micro[45] — that has a core business interest in the security of a wide range of digital technologies produced by other vendors. As Mårten Mickos, CEO of HackerOne, observed in his interview with AJL, initiatives like Google's Project Zero are "pushing the industry to accept … that you have to fix [vulnerabilities] without delay … [but] it probably wouldn't work if it weren't such a respected company doing it."[46] There are a very limited number of institutions that have the broad-based trust across industry and within the security research community needed to make that kind of full disclosure model work, as well as the resources and business incentives to ensure fair compensation for researchers. Google and Trend Micro are in that sense exceptional (and have also recruited exceptional groups of researchers).

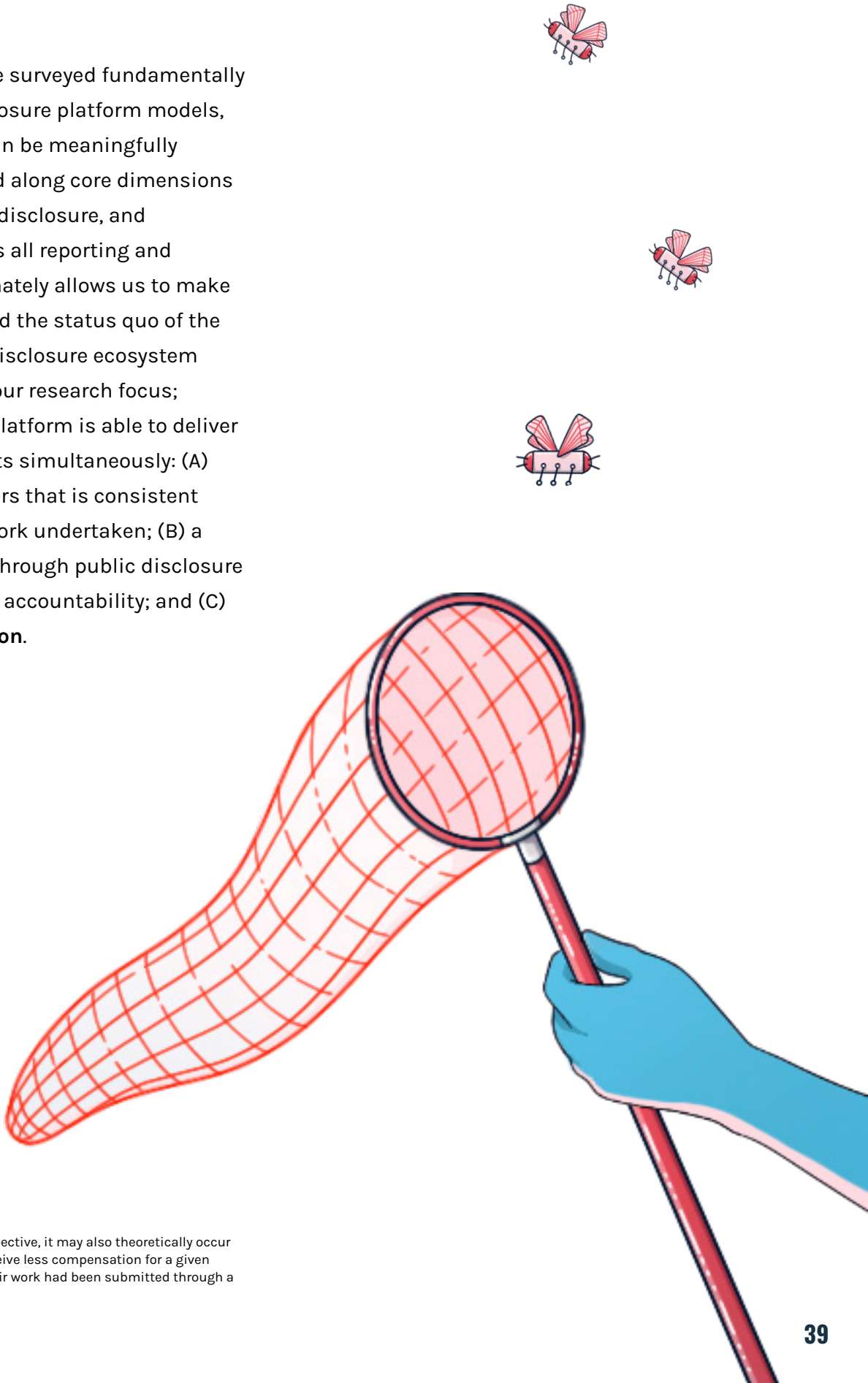
It is similarly worth noting that some 'closed' reporting and disclosure platforms, which tend to operate largely with a norm of non-disclosure and closely vet researchers for skill and trustworthiness before they are invited to participate, have increasingly come to look like or even explicitly brand themselves as penetration testing mechanisms (e.g., Cobalt.io), rather than platforms explicitly offering BBPs (Hansen 2016). This shift aligns with target organizations' interest in exercising a greater amount of control over authorized vulnerability research, such that security scrutiny can be most readily targeted to align with those organizations' objectives and priorities. While the labor conditions for researchers may be in some ways more favorable under these kinds of setups (i.e. they are paid for their hours worked, rather than their findings), that positive aspect — given the contractual relationship through which it is achieved — also implies a

45 In the case of ZDI, the platform has had several corporate benefactors since its founding, due to successive acquisitions, mergers, and the like.

46 A JL Interview: Mårten Mickos

trade-off on transparency in that researchers and pentesting reporting and disclosure platforms can rarely disclose discovered vulnerability information publicly.[47]

To summarize, while we have surveyed fundamentally different reporting and disclosure platform models, these various approaches can be meaningfully compared and distinguished along core dimensions of compensation, reporting, disclosure, and participation. Looking across all reporting and disclosure approaches ultimately allows us to make qualitative assertions around the status quo of the cybersecurity vulnerability disclosure ecosystem that are directly relevant to our research focus; namely that, at present, no platform is able to deliver on all of the following aspects simultaneously: (A) **compensation** for researchers that is consistent and commensurate to the work undertaken; (B) a guarantee of **transparency** through public disclosure sufficient to ensure ongoing accountability; and (C) open, accessible **participation**.



47 From a researcher compensation perspective, it may also theoretically occur that full-time, salaried researchers receive less compensation for a given "find" than what they would have if their work had been submitted through a BBP.

Figure 3: Reporting and Disclosure Platform Variations Relative to Reference Examples

PLATFORM NAME	HackerOne[48] (Reference Public BBP Platform)	Synack[49] (Reference Private BBP Platform)	Open Bug Bounty[50]	WooYun (乌云网)[51] (Forced to shut down in 2016 by Chinese authorities)	Trend Micro Zero Day Initiative (ZDI)[52]	Google Project Zero[53]	Cobalt.io[54]
REPORTING MODEL	Voluntary	Voluntary	Adversarial	Adversarial	Adversarial	Adversarial	Voluntary
TYPICAL DISCLOSURE MODEL	Limited / Non-Disclosure	Non-Disclosure	Delayed Full Disclosure (<90 days)	Delayed Full Disclosure (<45 days)	Delayed Full Disclosure (<120 days)	Delayed Full Disclosure (<120 days)	Non-Disclosure
PARTICIPATION MODEL	Open / Closed	Closed	Open	Open	Open	Closed	Closed
COMPENSATION MODEL	Bounties	Bounties	"Bounties (Not Guaranteed)"	"Bounties (Not Guaranteed)"	"Bounties + Bonuses"	Employment	Contract (Up-front payment)"

"Open" (public) models for adversarial research and reporting struggle to reliably and adequately compensate researchers while report quality is a significant issue.

"Closed" (private) models for adversarial research and reporting have only achieved success under corporate ownership, which risks limiting the scope of research.

The line between private BBPs and pentesting is blurring.

48 HackerOne is the largest BBP platform in operation by vulnerabilities reported and number of actively participating hackers. The company offers both public and private BBPs, as well as VDPs and crowdsourced pentesting services.

49 The Synack Red Team FAQ indicates that the company conducts report triage and determines bounty amounts, including for "Missions" that may include patch verification or security testing unrelated to vulnerabilities (e.g., checking for the use of common passwords) (Synack, n.d.).

50 Open Bug Bounty is a not-for-profit platform that offers report validation and facilitates initial contact between researchers and target organizations. Bounties are not guaranteed, and researchers are only able to report common web vulnerabilities (Open Bug Bounty, n.d.).

51 WooYun operated under a similar approach to Open Bug Bounty, with the platform attempting to engage organizations about which reports had been submitted. Most vulnerabilities reported through the platform did not lead to the payment of monetary bounties (Zhao 2016).

52 According to the ZDI FAQ, once a researcher accepts a reward offer, they are not authorized to directly disclose their findings (Trend Micro, n.d.). Rather, Trend Micro works with the affected vendor to establish a timeline for disclosure of up to 120 days (for responsive vendors) (Trend Micro, n.d.).

53 The Project Zero team searches for vulnerabilities in the most widely used software and hardware and reports their findings to the relevant manufacturer (Google, n.d.).

54 Around 2017, Cobalt.io began transitioning towards offering 'pentesting-as-a-service,' rather than BBPs, premised on the basis of a better ratio of valid to invalid reports and lower overall costs for target organizations (Hansen, 2017). Under this model, researchers are compensated on a contractual basis for services performed, rather than just based on vulnerabilities that they identify while testing (i.e. bounties) (Cobalt, n.d.).

IV. KEY RESEARCH TAKEAWAYS

In exploring our core research questions, we identified a number of key takeaways that shape our overall understanding of the applicability of BBPs and related models to a broader set of socio-technical harms, notably algorithmic harms.

KEY TAKEAWAY

1. PREPARE TO INCLUDE SOCIO-TECHNICAL CONCERNS

A handful of players in the BBP ecosystem have been slowly expanding their current programs to include socio-technical issues. This hasn't happened in a structured way, and no clear best practices have emerged yet, but the trend is likely to continue accelerating. As with any bug bounty, it will be essential for organizations to establish the 'digestive systems' necessary to meaningfully process and address the concerns identified through these mechanisms.

"BEST PRACTICES, INDUSTRY STANDARDS, TECHNICAL STANDARDS, [AND] INTERNATIONAL STANDARDS WILL PLAY A VERY IMPORTANT ROLE … TO ADVANCE SOCIAL UNDERSTANDING … AROUND DISCLOSURE [OF ALGORITHMIC HARMS], … AROUND ACCESS AND ACCESS TO WHAT, [AND] THE TYPE OF AUDIT. … THIS IS WHERE WE ARE GOING TO NEED TO ADVANCE OUR UNDERSTANDING OF THE ISSUE AS PART OF THE BROADER AI GOVERNANCE CONVERSATION."

— Amit Elazari Bar On, Director, Global Cybersecurity Policy, Intel Corporation[55]

When we started this project, the question of how quickly, if at all, BBP-like mechanisms would emerge for algorithmic harms or other socio-technical issues seemed wide open. The idea of "bias and safety bounties for AI systems" among various other prospective mechanisms was briefly explored through a workshop report released in early 2020, but no systematic investment in such mechanisms looked to have been made at that time (Brundage et al. 2020).[56] However, with closer scrutiny, including of subsequent developments in this space, it appears that such a progression is not only already well-underway, but accelerating. This trend should motivate more research into the design levers and tradeoffs that can be leveraged by these kinds of programs, along with broader consideration of how to get them to succeed.

55 AJL Interview: Amit Elazari Bar On

56 Note: Some reporting around this Brundage et al. (2020) claims that one of that paper's co-authors "initially suggested the idea of bias bounties for AI" in 2018 (Johnson 2020), however, that assertion is inaccurate, notably as prior internal work on bias bounties was conducted at Google by one of this report's co-author as far back as 2016.

"[A] LITTLE BIT OF TIME FROM HERE AND THERE CAN CREATE THINGS LIKE OPEN SOURCE SOFTWARE, WIKIPEDIA, AND STUFF BY THESE SMALL CONTRIBUTIONS ACROSS THE LONG TAIL … THE BIG QUESTION I HAVE ABOUT … THE IDEA OF … BUG BOUNTIES [IN] OTHER DOMAINS IS WHETHER THERE IS A COGNITIVE SURPLUS EFFECT THAT CAN DO THIS, OR WHETHER IT'S SOMETHING THAT [HAS TO BE] DRIVEN BY A SMALL NUMBER OF PEOPLE WITH DEEP SUBJECT MATTER EXPERTISE, WHO MAY NOMINALLY HAVE JOBS AND MAY NOT FEEL LIKE APPLYING THIS EXPERTISE IN THEIR [OFF] HOURS … AND ALSO, THE OTHER QUESTION IS WHETHER THEY CAN RELIABLY DO THIS WORK FROM AFAR, WITHOUT CONTEXT [OF] WORKING IN [THE RELEVANT] ORGANIZATION."

— Dino Dai Zovi, Security Researcher[57]

Two of the earliest programs to emerge were Google and Facebook's "data abuse" BBPs, created following the Cambridge Analytica scandal. These sit within the conceptual penumbra of algorithmic harm, and reward individuals for accurately reporting and providing evidence revealing instances of data misuse (e.g., the sale, resale, or repurposing of user data by third party affiliates, such as app developers, beyond what is authorized by Google / Facebook), while avoiding the more challenging-to-address harms that intended and authorized uses of such data may induce (Bacchus, Porst, and Mutchler 2019; Greene 2018). From the creation of Google's abuse-oriented program through to June 2021, the company reported receiving and accepting over 1,000 submissions (Hupa, Henson, and Straka 2021).

In addition, since 2015, Google has offered research grants to top-performing contributors to its cybersecurity BBP, tied to research activities including testing new products and features, scrutinizing sensitive Google services, and validating security fixes (Google, 2021). Researchers are expected to conduct work scoped to the terms of the grant, but are not required to identify flaws in exchange for that compensation.[58] In June 2021, the company further expanded the scope of this program to cover "[p]roduct and feature abuse," inviting top contributors from the company's above-referenced socio-technical BBP to apply for grants up to $3,133 to conduct research on those topics (Hupa, Henson, and Straka 2021; Google 2021).

"USUALLY, [COMPANIES] START WITH TECHNICAL SECURITY VULNERABILITIES AND THEN THEY EXPAND SCOPE OUT. IN A FEW CASES, THE SCOPE EXPANDS SO MUCH THAT IT FORKS INTO TWO SEPARATE POLICIES OR TWO SEPARATE PROGRAMS. BUT MOST OF THE TIME, IT'S [THAT] THEIR CORE PROGRAM STARTS GRADUALLY EXPANDING SCOPE OVER TIME."

— Alex Rice, CTO, HackerOne[59]

57 AJL Interview: Dino Dai Zovi

58 This move by Google towards grant-based funding, could also be interpreted as a recognition of and response to the income instability that comes with BBP work. In April 2020, the company even announced a new addition to its "Vulnerability Research Grants" to support contributors to Google's BBPs during COVID-19 (Jupa 2020).

59 AJL Interview: Alex Rice

Rockstar Games, meanwhile, offers a bug bounty for anyone who can prove that a player has been incorrectly banned from its online gaming platform for cheating, a process which can involve algorithmic decision-making (e.g., to identify unusual patterns of in-game behavior) (Rockstar Games 2021).

"[S]OME OF THESE COMPANIES MIGHT BE NERVOUS ABOUT SUBMITTING THEIR ALGORITHMS FOR SCRUTINY IF THEY'RE NOT CONFIDENT IN HOW THEY WORK. FOR THEM, THERE'S A POTENTIAL FOR GREAT PUBLIC EMBARRASSMENT IF … [THEY]'VE BEEN USING THIS ALGORITHM LIVE ON [THEIR] PLATFORM, AND … HAD A BUNCH OF RESEARCHERS TAKE A LOOK AT IT, AND THEY JUST TORE IT TO SHREDS."

— Marcia Hofmann, Digital Rights Lawyer[60]

Most recently, Corellium, which provides emulation tools for researchers, application testers, and journalists seeking to communicate anonymously and securely, established an initiative providing a limited number of grants to researchers aiming to reveal ways in which mobile software vendors were failing to live up to their promises on privacy, as well as security (Corellium 2021). Again, sitting somewhat adjacent to the conceptual core of algorithmic harm, Corellium's approach nonetheless constitutes an interesting proof of concept for supporting adversarial harms research, although is limited in its scope and possible impact with respect to both funding ($5,000 per grantee) and participation (only three grant openings available).

Twitter's August 2021 announcement of a 'bias bounty challenge' represented a new and important milestone in these developments, which we explore in further detail below in section 5, Twitter's Algorithmic Bias Bounty Challenge.

"I DO THINK THE MODEL IN GENERAL IS APPLICABLE [TO THE CONTEXT OF ALGORITHMS]."

— Mårten Mickos, CEO, HackerOne[61]

These examples all illustrate the feasibility of stretching BBP-like mechanisms to cover a wider range of socio-technical issues. Much work remains to properly contextualize these approaches within a robust conceptual and design framework that fully reckons with the trade-offs associated with different programmatic and institutional approaches.

60 AJL Interview: Marcia Hofmann

61 AJL Interview: Mårten Mickos

2. LOOK ACROSS THE LIFECYCLE

The allure of BBPs can often obscure the fact that they are just one mechanism for enhancing cybersecurity that fits within a broader, ongoing arc of field maturation and organizational development. A similar 'technology product lifecycle' lens to that which exists in security (as the 'secure development lifecycle') should be applied to algorithmic harms.

"[THE] BEST BUG BOUNTY CUSTOMERS WE HAVE, THEY TAKE EVERY BUG, THEY GO BACK TO SOFTWARE DEVELOPMENT AND SAY, 'WHAT CAN WE LEARN FROM THIS BUG ON A DEEPER LEVEL?' NOT JUST FIXING [THAT VULNERABILITY], BUT FIXING THE WAY [THEY] DEVELOP SOFTWARE."

— Mårten Mickos, CEO, HackerOne[62]

BBPs, and their ability to impact the dimensions we care about — creating a community of practice, advancing the state of the field and the practice, and increasing transparency and accountability across the industry — are most efficient as one step in a wider and longer process of field development and organizational maturation, plus adoption of best practices and standards for preventing and addressing vulnerabilities.

Successfully applying the BBP model requires a certain foundation: specifically, pre-existing communities of practice, a mature field with well-defined problems, and a scalable, well-resourced system for translating the content of reports into actions that address the problems described in those reports. A clear understanding is needed of who can contribute to the surfacing of particular issues (e.g., academic / technical researchers, impacted communities, journalists, etc.), and how those contributions should be channelled and leveraged. Problems need to be defined and scoped in such a way that reporting can be somewhat standardized (refer to Design Lessons 11 and 12). Overall, target organizations need to have appropriate and well-resourced "internal digestive system[s] for vulnerabilities," a metaphor deployed by Katie Moussouris to describe the need for processes for ingesting, responding to, and rectifying whatever is to be surfaced and reported through a BBP (Moussouris, 2021b). As has been the case in the cybersecurity context, considering how best to identify and mitigate risks across the full technology product lifecycle is essential to the development of robust 'digestive systems' (refer to Tackling Cybersecurity Vulnerabilities and Algorithmic Harms for additional details).

62 AJL Interview: Mårten Mickos

“THE WORLD HAS AGREED THAT JOURNALISTS AND AUTHORS NEED PROOFREADERS ALWAYS, BECAUSE YOU DON’T SEE YOUR OWN TYPOS. NO BIG DEAL. BOOKKEEPING IS GREAT, BUT WE ALWAYS NEED AUDITORS, OTHERWISE YOU CAN’T TRUST BOOKKEEPING. EVERYBODY IS FINE WITH IT, NOBODY SEES IT AS A SIGN THAT BOOKKEEPING AS A PRACTICE WOULD BE COMPLETELY FLAWED. WE HAVE TO HAVE CHECKS AND BALANCES.”

— Mårten Mickos, CEO, HackerOne[63]

For the growing number of organizations that already have robust BBPs in place, extending these existing cybersecurity-oriented programs to meaningfully cover algorithmic harms is harder than simply adding algorithmic systems within their existing program scopes. This is because the resolution of algorithmic harms needs to proceed through different ‘digestive systems’ within those organizations, as described above. Internally, organizations’ approaches to these issues need to be documented in robust, enforceable policies and implemented by appropriate teams to facilitate analysis and mitigation, with any broader lessons learned from reported issues incorporated into relevant design and production processes. Simply adding algorithmic harms to the scope of a company’s existing security BBP without the instantiation of new internal processes and additional resourcing for the teams responsible for developing and maintaining those algorithmic systems is unlikely to meaningfully reduce those systems’ harmful effects.

All experts and practitioners we interviewed in the course of this work agreed that BBPs were not a comprehensive solution for vulnerability management, let alone for cybersecurity writ large, and warned against the widespread preconception that BBPs could be quickly and easily stood up as a way for any given organization to ‘fix’ its security issues. Hence, BBP-like mechanisms may not be immediately impactful as a standalone tool in nascent fields, for young organizations, or in the absence of complementary development processes with widely used best practices and standards. Rather, BBPs are best thought of as one possible avenue for redress when situated within a broader, well-resourced, and mature product development lifecycle.

“WE OFTEN HAVE INFORMATION SECURITY TEAMS LAUNCHING A BOUNTY PROGRAM, AND IF AN INFORMATION SECURITY TEAM LAUNCHES A BOUNTY PROGRAM, THEY’RE ONLY GOING TO INCENTIVIZE THE THINGS THAT THEY’RE IN A POSITION TO DO SOMETHING ABOUT. SO IT’S NOT LIKE ‘ACME INC.’ LAUNCHES A BOUNTY PROGRAM; YOU HAVE TO PEEL IT BACK … TO [ASK] WHICH TEAM LAUNCHED THAT PROGRAM [AND] WHAT TYPE OF FEEDBACK ARE THEY IN A POSITION TO DO SOMETHING ABOUT?”

— Alex Rice, CTO, HackerOne[64]

63 AJL Interview: Mårten Mickos

64 AJL Interview: Alex Rice

3. NURTURE THE COMMUNITY OF PRACTICE

Bug bounty platforms play both a direct and indirect role in nurturing communities of practice, providing educational materials and tools, and motivating community members to independently develop and share resources. We believe there is a strong need for similar, well-curated, accessible resources to help nurture the algorithmic harms field.

"[HACKATHONS ARE] INTERESTING, BECAUSE, OF COURSE, THEY'RE FOR THE HACKERS, BUT THEN THEY'RE ALSO A MARKETING OPPORTUNITY FOR THE COMPANIES … WHERE RATHER THAN JUST BEING A DAY OF HACKING, [THE COMPANIES] GIVE YOU ACCESS TO THE … SCOPE, MAYBE THREE WEEKS BEFORE, AND THEN YOU'VE THREE WEEKS TO DO ALL OF YOUR HACKING AND THEN THEY JUST PAY THE BOUNTIES ON THE DAY OF THE EVENT … IT'S BETTER FOR THE PLATFORMS AND COMPANIES, SINCE THERE ARE MORE BUGS AND MORE BOUNTIES, BUT THAT DAY ITSELF IS PROBABLY LESS ACTUAL HACKING, AND MORE … [ABOUT] WATCHING THE RESULTS COME IN."

— Jack Cable, Security Researcher[65]

Independent researchers' contributions to security through formal mechanisms, such as BBPs, are increasingly recognized, thanks in-part to heightened awareness of cybersecurity across organizations and national boundaries. However, it is important to note that, historically, the curation and sharing of vulnerability information emerged first through informal, pre-internet communities (for example, via bulletin board systems) which, by the 1990s, had shifted to online forums. Notable among these was Bugtraq, whose significant impact on norms-setting and community-building across the security research ecosystem is detailed by Matt Goerzen and Gabriella Coleman (Forthcoming). For instance, these researchers note that Bugtraq enabled "[participants with] emails from institutional domains like @nasa.gov., @mitre.org, and @ufl.edu [to be] in dialogue with emails originating from private individuals — hackers and independent security researchers — with edgily named domains like @crimelab.com, @panix.com, and @dis.org." BBP platforms have subsequently played a supporting role in furthering this transformation by guiding organizations through the benefits of vulnerability reporting mechanisms

65 AJL Interview: Jack Cable

and creating opportunities for security researchers to collaborate, as well as acquire and build relevant skills.

"[IF] RESEARCHERS REALLY DO START TO FEEL A PART OF THE BROADER TEAM … THAT'S HOW YOU'RE GOING TO GET [THAT] GROUP … TO CONTINUE TO WANT TO DO THIS FOR THE LONG RUN."

— Lisa Wiswell Coe, Fmr. Program Manager, Hack the Pentagon[66]

Overall, BBPs tend to create, encourage, and curate materials and resources that can be useful in fostering a broader community of practice for security research. For example, public bug reports, when made available on BBP platforms, serve as a key learning resource. Zhao suggests that past vulnerability reports published on the platforms enable hackers "to learn valuable technical insights and skills from others' findings" (2016, 84). This assertion is supported by analysis of data on users that followed particular vulnerability reports on Wooyun, which shows that high-complexity vulnerabilities are followed by large numbers of participants, in turn suggesting they are recognized as containing valuable, teachable insights. However, even some of the less technically complex reports analyzed by Zhao also attracted significant attention, indicating that they too can be a useful baseline learning tool (e.g., for entry-level researchers) (Ibid., 86).

"I THINK THAT THE RECRUITING AND RETAINING EFFORTS HAVE REALLY HELPED THIS COMMUNITY PROLIFERATE … IT RELIES ON FINDING RESEARCHERS WITH SPECIFIC SKILL SETS AND … TAGGING THEM AS SUCH … LIKE THE 'ICS DATA GUY' OR … 'THE EMBEDDED IOT [DEVICE] GUY' … [SO THEN] PEOPLE FALL INTO THESE LANES WHERE THEY KNOW THAT THEY CAN CONTRIBUTE."

— Lisa Wiswell Coe, Fmr. Program Manager, Hack the Pentagon[67]

BBP platforms also offer learning resources and opportunities to help identify and develop talent (e.g., capture-the-flag, "bring a friend" for hackathons, etc.),[68] including formal educational tools to help individuals new to security research develop a baseline level of technical skill. HackerOne, by way of illustration, includes in its "Hacker101" video series an overview of the prerequisite knowledge for security research; how to identify, exploit, and remediate the most widespread web security vulnerabilities; how to write a strong bug report; and how to make effective use of one of the most popular security research tools, Burp Suite (HackerOne 2018).[69] HackerOne also offers a designated point of contact for security researchers to reach out with questions about the course material. A related offering popular with bug bounty platforms, ostensibly for skill development, is the Capture the Flag (CTF) format. Security researchers can work their way through a series of challenges where they leverage security

66 AJL Interview: Lisa Wiswell Coe

67 AJL Interview: Lisa Wiswell Coe

68 AJL Interview: Mårten Mickos

69 Bugcrowd makes a similar educational offering available, also free of charge; "Bugcrowd University" is arranged into modules with videos, slide decks, and practical "labs" hosted on GitHub (Bugcrowd, n.d.). YesWeHack, a BBP platform based in Europe, also offers training environments for specific vulnerability types through its "Dojo" (YesWeHack 2020) .

flaws to locate a piece of code (the "flag") contained within a mocked-up digital environment. While these CTFs may have some value in allowing novice hackers to practice their skills in a contained, risk-free environment, the primary goal for platforms offering these CTFs is once again recruitment-centric. HackerOne CEO Marten Mickos describes the arrangement as follows:

> *Everybody who signs up — we can see how quickly they find [the flag]. It's a little bit like an orienteering task. And when they find it … that's a good sign. The faster they find it, the better … and we can thereby source … out, digitally, in a scalable way … [hackers who are] very good on the CTF … and then we invite … [them] into a [private] program and we see how they perform.*[70]

Beyond opening opportunities on these platforms themselves, participation in BBPs can also open up career opportunities in the field, allowing researchers to create a portfolio of work that can later be leveraged on the job market. Adjacent to the major platforms, are offerings such as the free hacking education program "PortSwigger Web Security Academy," which is produced and maintained by the developer of the aforementioned Burp Suite (PortSwigger, n.d.).

Although the efficacy of these more formalized hacking education programs is unclear, such resources may be usefully emulated by organizations operating in the algorithmic harms space to expand access to relevant skills and tooling. This kind of programming could include practical lessons from technical experts on topics such as how to identify underlying training libraries used in algorithmic systems or how to sample and evaluate a classifier's output given different models, as well as meta-education in how to further educate oneself about fairness, bias, harms, and disparate impact in algorithmic systems. It could also usefully entail the identification and curation of key resources including benchmark datasets, code libraries useful for measuring fairness in algorithmic systems, and open-source research tools, such that those resources can be more easily accessed by newcomers.

"[THE] COMMUNITY-BUILDING ACTIVITY WILL HAPPEN WITH OR WITHOUT SOMETHING LIKE HACKERONE. BUT WHEN THERE IS A STEWARD … COORDINATING THE WORK BETWEEN COMPANIES AND HACKERS, THERE IS … THE OPPORTUNITY TO ESTABLISH PRINCIPLES AND RULES FOR ALL TO FOLLOW."

— Mårten Mickos, CEO, HackerOne[71]

In addition to direct educational offerings from reporting and disclosure platforms and adjacent businesses, a thriving bug bounty community exists on social media platforms where researchers offer guidance and tutorials. Popular BBP channels

70 AJL Interview: Mårten Mickos

71 AJL Interview: Mårten Mickos

on YouTube produce a range of content including instructional videos focused on particular technical skills or vulnerability types, overviews of high-profile vulnerabilities surfaced by security researchers, and broader guidance on how to approach learning and developing relevant technical and organizational skills. These often go into a great degree of depth and pull from these bug hunters' own experiences to extend the standard informational offerings provided by BBP programs and other formal organizations.[72] Through interviews with various security researchers, Ryan Ellis and Yuan Stevens revealed that "[n]umerous people … run blogs or use social media to share their bug bounty work … [including] in the hope of … 'getting knowledge out there, getting recognized for that, and helping other people learn.'"[73]

"A NICE THING ABOUT THE COMMUNITY IS THAT … MOST OF THE TOOLS [AND] METHODS OUT THERE ARE PUBLIC. SO PEOPLE WILL OPEN-SOURCE THEIR TOOLS [AND] WRITE ABOUT [THEM]. … I'D SAY THERE IS A … SPIRIT OF … SHARING WHAT YOU FIND [AND] WHAT YOU MAKE OUT INTO THE OPEN."

— Jack Cable, Security Researcher[74]

However, there are some notable tensions between the positive, community-enabling aspects of BBPs and other dynamics that their design encourages, which may be detrimental to community-building objectives. As Stevens and Ellis noted to us, despite some programs enabling researchers to split their bounties, the 'winner-take-all' dynamics of these markets don't actively encourage collaboration,[75] since the size of each bounty is generally contingent on the security impact of the discovered vulnerability, rather than directly on the amount of work required to unearth it, and discovery of that vulnerability is only rewarded once. In addition, as noted by Meredith Whittaker, BBPs and other "technocratic remedies … and fairness fixits … cast elite engineers as the arbiters of 'bias,' while structurally excluding scholars and advocates … whose focus on the racialized power asymmetries and political economy of AI are essential for understanding and addressing AI harms" (Whittaker, 2021). An intermediary for algorithmic harms reports should thus look to cultivate and motivate greater collaboration among more diverse participants, including advocates and socio-technical scholars whose role has been and must remain indispensable to awareness and redress of algorithmic harms. For example, they could structure reward bonuses to encourage multidisciplinary, participatory, collaborative work, which may be especially useful for motivating the discovery of complex issues requiring more work and a greater variety of skills to identify and examine. A comparable reimagining of current cybersecurity reporting and disclosure models to focus on encouraging growth and diversification of participant communities would also be an interesting avenue for future research.

72 For example, Nahamsec's video on reconnaissance — a process where hackers leverage various tools to uncover how a particular digital asset is structured and what underlying technologies it uses — details how to approach that task for BBPs of varying scopes. Another YouTuber, STÖK, has almost 400,000 views on a video relating his experience and advice about getting started as a security researcher.

73 Interview with Jack Cable in Ellis and Stevens (Forthcoming)

74 AJL Interview: Jack Cable

75 Notes from conversation between the authors of this paper, Ryan Ellis, and Yuan Stevens (September 2021).

KEY TAKEAWAY

4. INTENTIONALLY DEVELOP A DIVERSE COMMUNITY

A small number of bug bounty platforms dominate the vulnerability disclosure ecosystem, leading to concerns about the commodification of security research and lack of diversity among these platforms' contributors. Deploying bounty programs successfully for algorithmic harms requires serious effort to recruit and retain diverse communities of researchers and affected individuals.

"THE VALUE-ADD OF THESE PROGRAMS IS THE LONG TAIL, IT IS IN THE DIVERSITY. WE COULD RUN OUR COMMERCIAL BBP BUSINESS WITH THE TOP ONE-PERCENT OF THE COMMUNITY. YOU COULD PICK EVEN LESS THAN THAT. YOU COULD PICK 1,000 HACKERS AND RUN THE COMMERCIAL BUSINESS AT HACKERONE, AND SOME FOLKS DO, BUT YOU'RE MISSING OUT ON A MASSIVE PART OF THE FEEDBACK CYCLE BY NOT ENABLING AND ALLOWING THE LONG TAIL."

— Alex Rice, CTO, HackerOne[76]

Industry leaders such as Katie Moussouris (Marino 2020) and researchers such as Ryan Ellis and Yuan Stevens have pointed to the role that VC-funded bounty platforms have played in "turning bugs into property and hackers into gig workers" (Ellis and Stevens Forthcoming). Ellis and Stevens' review of marketing materials produced by these platforms highlights that they explicitly paint their BBP offerings as a "cheaper alternative to either in-house work or expensive pen test contracts" (2021). However, for most workers, BBP income is unreliable and incomplete — a problem often found elsewhere in the 'gig economy' — meaning that some other source of economic stability is required for even talented security researchers to commit significant time to independent security research. BBPs and BBP platforms do not offer typical employment benefits (e.g., healthcare) for security researchers, while bounty-derived income is earned on a highly uncertain, periodic basis with reward amounts calculated based on imperfect metrics (e.g., the Common Vulnerability Scoring System, or CVSS).[77] Moreover, the amount of work required to find the 'next' vulnerability is hard to anticipate, only the first discovery of a vulnerability is rewarded, and, as Charlie Miller puts it, "vulnerability information is a

76 AJL Interview: Alex Rice

77 For an exploration of the methodological, operational, and practical flaws of CVSS, refer to Spring et al. 2018 and 2021.

time-sensitive commodity" (Miller 2007, 2).[78]

Faced with this vein of criticism, defenders of BBPs stress that comparing their marketplaces with other 'gig economy' platforms such as Uber, InstaCart or Amazon Mechanical Turk misrepresents the nature of the work performed and the context around it: security researchers participating in BBP platforms bring differentiated skills, and as a result their experiences as workers on these platforms tend to vary greatly.[79] Researchers analyzing datasets of bug bounty submissions from both HackerOne and Facebook's programs have demonstrated that more technically skilled researchers not only earn a disproportionate volume of total bounties awarded, but are courted by the various platforms and programs who compete to attract and retain the interest of these researchers (Ellis, Huang, Siegel, Moussouris, and Houghton 2017, 134). This in turns gives these 'elite' security researchers a different power relationship with platforms than the great majority of participants, which, for instance, enables them to contest decisions made by platforms around the validity or appropriate compensation level for reports submitted (Ellis and Stevens Forthcoming).

In addition to concerns around worker empowerment and fairness, most bug bounty platforms struggle with creating and retaining diverse and inclusive communities of practice. In annual reports produced by one of the major BBP platforms, Bugcrowd, ninety-four percent of researchers surveyed in 2020 identified as male (Bugcrowd 2020, 13), a two-percent increase since 2019 (Bugcrowd 2019, 5), with the remaining six percent identified as female — a dismal gender gap, along with a statistical absence or erasure of non-binary people in this survey data. Data on the racial and ethnic diversity of Bugcrowd's research community suggests that the largest ethnic group is "Asian or Pacific Islander," at forty-six percent of the total surveyed population (Bugcrowd 2020, 12). However, this figure speaks more to the platform's role in offshoring security work than it does to any efforts to address demographic imbalances across the platform's community.[80] Meanwhile, "African American" and "Hispanic" researchers each constitute only three percent of participants on the platform, and intersectional analysis of the Bugcrowd community is entirely missing from these reports. Despite the sparsity and superficiality of this data, it is clear that more work is needed to understand and address this troubling lack of diversity both within the security research community and across the field more widely, which otherwise will continue to impede organizational and societal efforts to improve cybersecurity (Stewart 2020).

"[BBP WORK] IS LIKE BEING AN UBER DRIVER, ONLY YOU [AND] THOUSANDS OF OTHER DRIVERS ALL DRIVE TO THE AIRPORT, TO FIND THAT ONLY [ONE] OF YOU WAS LUCKY ENOUGH TO HAVE A PAYING PASSENGER."

— Katie Moussouris, CEO, Luta Security (2021a)

78 An interesting characteristic of vulnerability information is that its value can rapidly go from extremely high to near zero, since, like other informational goods, it is only valuable when it is not widely known. As soon as new vulnerability information is publicized, even before a patch or other effective mitigation has necessarily been released, that information swiftly becomes worthless (Miller 2007, 2-3).

79 AJL Interview: Amit Elazari Bar On

80 Ninety-nine percent of total bounties paid out in the year prior to Bugcrowd's "Inside the Mind of a Hacker 2020" report came from organizations based in the U.S., South Korea, Australia, and Europe. Thirty-nine percent of this money went to researchers based just in India and Pakistan (Bugcrowd 2020, 9).

Much of the existing academic literature on BBPs tends to obscure these problems through a disproportionate focus on economic efficiency, narrow framings of researcher 'diversity' (e.g., differentiation solely by skill level or technical competencies, without serious consideration of demographic factors), and similarly limited consideration of the security impacts that result. However, things are beginning to change. Ellis and Stevens foreground the labor issues of BBPs in their recent research conducted with Data & Society (Forthcoming). Community-driven initiatives, notably #ShareTheMicInCyber, are surfacing and working to address "issues stemming from systemic racism in cybersecurity" (Stewart and Zabierek, n.d.). Established academics continue to compile data on the real-world experiences of security researchers through initiatives including the "Better Bug Bounties" project (Elazari et al., n.d.).

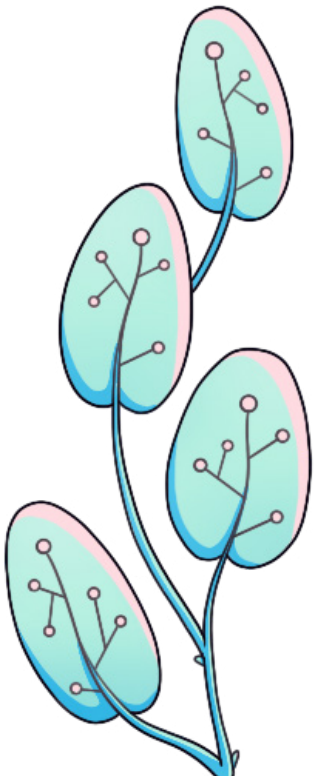
In practice, addressing these concerns will require platforms that host BBPs to hold themselves to account with respect to both transparency and action, for example, in establishing meaningful diversity goals for their communities and analyzing and publicizing — in consistent format and with relevant context — available data on their progress.

"THE MAJORITY OF VALUE THAT PROGRAMS ARE GETTING DOES COME FROM A VERY SMALL NUMBER OF HACKERS, AT LEAST RELATIVE [TO] WHAT THE PLATFORMS ADVERTISE."

— Jack Cable, Security Researcher[81]

"THE PLATFORMS ARE TRYING TO EXPAND THEIR COMMUNITIES … [AND] EDUCATE PEOPLE AND BROADEN THE SCOPE OF WHO'S PARTICIPATING … BUT EVEN THEN … IT REMAINS THAT THERE'S A RELATIVELY SMALL NUMBER OF PEOPLE WHO FIND MOST OF THE VULNERABILITIES [AND] THEY'RE NOT VERY DIVERSE."

— Jack Cable, Security Researcher[82]



81 AJL Interview: Jack Cable

82 AJL Interview: Jack Cable

5. FOSTER AND PROTECT PARTICIPATORY, ADVERSARIAL RESEARCH

The vulnerability disclosure space continues to struggle with participatory models for compensated, safe adversarial research. Unfortunately, this is what is most needed in the algorithmic harms space.

"[THE] BEST THING THAT THE GOVERNMENT COULD DO TODAY TO STRENGTHEN CYBERSECURITY AND THE RESILIENCY OF TECHNOLOGY IS TO NORMALIZE DISCLOSURE. FORCE THAT INFORMATION SHARING, AND THAT LEARNING AND COLLABORATION IN A SHARED WAY. … AND WHEN YOU LOOK AT THINGS LIKE THE AUTOMOTIVE AND AIRLINE INDUSTRIES WITH THEIR NTSB-MANDATED TRANSPARENCY REPORTS, THE CARROT THERE IS, … PUBLISH YOUR VULNERABILITIES AND WE WON'T BRING ANY LAWSUITS, OR LIABILITY SUITS, OR ANY REGULATION AGAINST YOU FOR THINGS THAT YOU VOLUNTARILY DISCLOSE ACCORDING TO THIS FORUM. THAT'S THE TYPE OF CARROT THAT ONLY GOVERNMENT REGULATION CAN PROVIDE THAT CREATES THE FEEDBACK LOOPS THAT EVERYBODY BENEFITS FROM, THAT WE DON'T HAVE TODAY."

— Alex Rice, CTO, HackerOne[83]

Transparency in the cybersecurity space, as well as in the algorithmic harms context, is a means to an end. Its value comes in the assurance of accountability that it can provide (e.g., through market-based or regulatory mechanisms) and in the resolution of information asymmetries between customers and vendors with respect to the security of those vendors' various offerings. Unfortunately, our research surfaced that ultimately, both for organizations actively participating in the vulnerability disclosure ecosystem and those that still refuse such engagement, such transparency remains entirely optional for target organizations. Individual security researchers seeking financial compensation for their work and the avoidance of legal risk, as well as reporting and disclosure platforms that are financially dependent either directly[84] or indirectly[85] on target organizations, are ultimately and necessarily constrained by the interests of target organizations. As such, the prototypical mediating solution that has been adopted for compensated vulnerability disclosures (i.e. platform-based BBPs) can end up disempowering both the platform and the

83 AJL Interview: Alex Rice

84 Direct dependence, e.g., reporting and disclosure platforms' reliance on target organizations contracted as clients .

85 Indirect dependence, e.g., a reliance on investment from venture capitalists also heavily invested in target organizations.

researcher, while ensuring that target organizations can continue to suppress, to any extent desired, the publicization of information regarding security weaknesses in their products and services.

"I THINK THAT IDEALLY WE'D HAVE A SITUATION WHERE BUG BOUNTIES HAVE TERMS [SUCH] THAT YOU CAN TALK ABOUT … VULNERABILITIES AFTER A SET … TIME, LIKE STANDARD COORDINATED VULNERABILITY DISCLOSURE PROGRAMS DO."

— Jack Cable, Security Researcher[86]

For the vast majority of prototypical BBPs, assurances of legal safe harbor and the promise of compensation are contingent on adherence to target organizations' discretion regarding when, what, and to whom researchers can further disclose their findings. By way of illustration, scrutinizing publicly available program terms on HackerOne reveals that only a small fraction of those that offer monetary rewards to researchers — i.e. BBPs, as opposed to VDPs — guarantee that researchers can disclose their findings on a pre-established timeline.[87] Furthermore, target organizations can unilaterally determine reports to be out-of-scope, in light of either the vulnerable system itself or the nature of a given vulnerability falling outside the organization's area of interest. In such cases, researchers are still often bound to those same limitations on public disclosure, even though their report has been dismissed as unwanted by the target organization. Even if an organization does ultimately take action in response to a researcher's report, that vendor or operator may take months or even years to fully address the identified flaw. In the meantime, customers are left unprotected and unaware, both of individual flaws and of their impact on the organization's overall security, since the researcher who reported the vulnerability is unable to publicize their findings without loss of compensation and risk of legal retaliation.

"RESEARCH SHOWS US THAT [CLEAR] EXPECTATIONS AROUND COMMUNICATIONS, INCLUDING AROUND DISCLOSURE, IS THE NUMBER ONE PRIORITY."

— Amit Elazari Bar On, Director, Global Cybersecurity Policy, Intel Corporation[88]

Moreover, security researchers deemed, even unfairly, to have violated a particular program's terms and conditions by publicly disclosing their findings without the consent of the vendor or operator may be banned from submitting further vulnerabilities. For example, in 2019, security researcher Vasily Kravets was briefly banned from using the HackerOne platform after he unsuccessfully attempted to submit a second, serious vulnerability to the BBP of gaming platform Valve, following the wrongful rejection of an earlier submission, and released

86 AJL Interview: Jack Cable

87 Authors' analysis of publicly available program terms on HackerOne as of February 2nd, 2021 for 338 programs, of which 178 offered monetary bounties. Of those 178, 51 programs (29%) explicitly authorized researchers to disclose their findings either on a universal timeline (e.g., 180 days post-submission) or on a patch-contingent timeline (e.g., 30 days post-patch). Included in this group are 22 programs (12% of the total, or 43% of the aforementioned group) that adhere to HackerOne's VDP terms, which include a contingency for disclosure by the researcher after 180 days, but also provide significant flexibility to target organizations with respect to deadline extensions and report redactions (HackerOne 2019b). As such, our findings should be interpreted as treating target organizations' affordance of freedom-to-disclose with a great deal of generosity. In addition to these findings for HackerOne's public programs, as of August 2019, 80% of all programs on HackerOne were private (HackerOne 2019c), and universally subject to "strict non-disclosure by default" (HackerOne 2019b).

88 AJL Interview: Amit Elazari Bar On

those details publicly in order to compel action by the company (Rash 2019). The major platforms that dominate the BBP ecosystem are not set up for adversarial research, but nor do they pretend to be, as HackerOne CEO Mårten Mickos notes: "We're not in the business of whistleblowing. We don't report on other people's problems. We report problems to people who have solicited that input, and that's an important distinction."[89,90]

"[T]HE COMPANY … HAS A LOT OF POWER TO DEFINE THE TERMS OF THE ENGAGEMENT … THEY'RE THE ONES WHO HOLD THE PURSE STRINGS; THEY DECIDE WHEN CONDITIONS ARE MET, AND WHAT THE PAYOUT IS GOING TO BE … I THINK THAT ALL THE PLAYERS IN THE ECOSYSTEM WHO WANT BOUNTIES TO BE SUCCESSFUL, ARE … INCENTIVIZED TO MEET THE[IR] REQUIREMENTS."

— Marcia Hofmann, Digital Rights Lawyer[91]

To the extent that compensated adversarial research does exist in the vulnerability disclosure space today, it is contingent on the financial independence of such initiatives from target organizations. The contrast between the disclosure approach of Google's Project Zero and Trend Micro's ZDI — publicization of discovered vulnerabilities no more than 120 days after the affected vendor has been notified — relative to the typical disclosure approach on a platform like HackerOne is useful in demonstrating where things have gone awry and why a solution to improving transparency with respect to cybersecurity vulnerabilities remains elusive.

"THE COMMON NARRATIVE IS, 'OH, THEY HATE TRANSPARENCY.' … THE REALITY IS, UNCOORDINATED DISCLOSURES … REPRESENT AN UNQUANTIFIABLE BUSINESS RISK TO MOST ENTERPRISES TODAY. … ONE OF THE MOST COMMON THINGS THAT WE HEAR FROM ORGANIZATIONS THAT ARE STRUGGLING WITH UNCOORDINATED DISCLOSURES IS, WHAT ABOUT THE COMPETITION? AND YOU HAVE TO BE SO FAR AHEAD OF YOUR COMPETITION TO BE THE FIRST ONE TO … PROACTIVELY DISCLOSE EVERY SECURITY VULNERABILITY … OR YOU HAVE TO COLLABORATE WITH ALL OF YOUR COMPETITION TO SAY, 'WE'RE ALL GOING TO DO THIS TOGETHER.'"

— Alex Rice, CTO, HackerOne[92]

Project Zero and ZDI, unlike most BBPs, disclose identified vulnerabilities on a predetermined timeline, even when doing so runs counter to the preferences of the relevant vendor or operator. However, as noted in Bug Bounty Programs 101, only Trend Micro and Google have proven able to sustain a model of adversarial disclosure while ensuring adequate compensation for researchers and high-quality report submissions, and have arguably

89 AJL Interview: Mårten Mickos

90 Notably, program terms for BBPs often include externally-facing lists of key systems / assets as part of the program scope. This characteristic could facilitate a currently missing baseline of visibility in the context of algorithmic systems about what systems are in use, and how they are being used. Again, though, companies' interests are unlikely to be consistently aligned with the identification of such systems — especially where doing so could induce further scrutiny."

91 AJL Interview: Marcia Hofmann

92 AJL Interview: Alex Rice

succeeded only as a result of their respective stature within the industry. That their economic and reputational interests align with a model of guaranteeing public disclosure of vulnerabilities largely affecting products on which the security of their own offerings is heavily dependent helps to explain why they would wish to adopt this approach. Their scale and reputation as global multinational corporations and, with that scale, the relative power and prestige that they hold within this disclosure ecosystem, explains why they have been able to adopt and maintain this model more effectively than the likes of WooYun or Open Bug Bounty. ZDI and Project Zero, while limited in scope to the most valuable and impactful security vulnerabilities, and not especially impactful as community-building institutions given their limited number or variety of participants, are exceptional in terms of compensating researchers under a (delayed) full disclosure research model, and thereby cementing accountability through an assurance of transparency.

"I THINK ONE OF THE BIG CHALLENGES WOULD BE GETTING THE COMPANIES … COMFORTABLE WITH THE IDEA [OF THIRD PARTIES SCRUTINIZING THEIR ALGORITHMS], AND SEEING THE VALUE OF THE PROCESS AND WHAT IT COULD DO FOR THEM. MAYBE IF THERE WERE SOME SORT OF REGULATORY SAFE HARBOR, THAT COULD BE SOMETHING THAT COULD BE HELPFUL, OR OTHER INCENTIVES."

— Marcia Hofmann, Digital Rights Lawyer[93]

"[IN THE 'INTERNET BUG BOUNTY' (IBB)], COMPANIES AND INSTITUTIONS … DONATE MONEY INTO THE FOUNDATION… TO PAY BOUNTIES ON BEHALF OF THESE FUNDED OPEN SOURCE PROJECTS … TO MAKE SURE THAT UNFUNDED OPEN SOURCE PROJECTS GET MONEY FROM SOMEWHERE TO PAY BOUNTIES, AND THEN WE HAVE GIVEN OUR PLATFORM AND ALL OUR SERVICE COMPLETELY FOR FREE TO THIS INITIATIVE. THE IBB GROUP HAS NO EXPENSES; ALL MONEY THAT COMES IN GOES TO HACKERS, AND WE OPERATE THE PLATFORM. AND WE JUST THINK THE WORLD WON'T GET SAFER IF WE DON'T DO THAT."

— Mårten Mickos, CEO, HackerOne[94]

This exceptional characteristic, and the financial independence that underpins it, stands out as a circumstance worthy of attention from would-be creators of BBPs for the algorithmic harms space. Financially independent intermediaries also rich in their own technical expertise may be better positioned to drive transparency and accountability in the algorithmic harms space than initiatives more akin to the prototypical in-house BBPs and platforms. Such institutions would need to be widely trusted and, hence, inclusive, transparent, and fair, as well as well-funded through mechanisms that would preserve their strategic independence — perhaps through support from governments, industry consortia, or philanthropists. Crucially, these intermediaries would need to be constructed

93 AJL Interview: Marcia Hofmann

94 AJL Interview: Mårten Mickos

to avoid undue influence from organizations that lack the credibility to lead or prominently fund such efforts, including Google, in light of its treatment of their own in-house algorithmic fairness researchers, and Facebook, given its tepid response to damning research on the harmful effects of its algorithmic products. Similarly, other corporate innovators in this space with business models dependent on inherently harmful algorithmic systems and products may need to be kept at arm's length from such initiatives.

"MY SENSE IS THAT PROBABLY THE BEST WAY TO TRY TO CREATE … SUCCESSFUL [DISCLOSURES] … IS TO REALLY HAVE … THE LAY OF THE LAND SET OUT SO THAT EXPECTATIONS ARE CLEAR … ON BOTH SIDES OF THE EQUATION. IT'S CLEAR YOU WILL GET A BOUNTY IF THESE CONDITIONS ARE MET, AND IF YOU DON'T MEET THOSE CONDITIONS, THEN YOU WON'T. I THINK ONE THING THAT A [PLATFORM] CAN PROVIDE IS THAT CERTAINTY, BUT ALSO [IT CAN PLAY] A FACILITATION ROLE, WHICH I THINK CREATES … TRUST."

— Marcia Hofmann, Digital Rights Lawyer[95]

Truly independent, well-resourced mediating institutions could empower impacted individuals to report experiences of harm while simultaneously supporting or coordinating the disclosure of research addressing those same concerns. These intermediaries could establish terms and conditions for researchers, reporters, and target organizations that would advance algorithmic justice by directing attention and resources, and create pipelines to funnel vetted reports and research to various, tailored audiences.[96] For example, those reports or research may be directed to journalists, public interest lawyers, or system administrators well-positioned to amplify people's stories, work towards legal redress, or make harm-reducing changes to system use or implementation, respectively. Matt Goerzen noted that this kind of transparency-enabling approach could also be of value to technical audiences outside of affected vendors and operators, in a manner that echo the early history of vulnerability disclosure, wherein "[some] of the biggest advocates and participants in early 1990s 'full disclosure' models were system administrators, who welcomed the ability to make their own local changes [to mitigate a newly identified flaw until] a vendor-approved patch [was made available]."[97]Over time, more of the information contained in those submissions could also be routed to vendors with a proven track record of addressing algorithmic harms throughout their product life cycles or operators with a demonstrated commitment to comprehensive redress, including fueling harm-reducing feedback loops for more robust and equitable development of systems in the future. In addition, these kinds of institutions could facilitate scrutiny of 'critical algorithmic infrastructure' (e.g., widely used datasets and open-source models) in a similar vein to how the 'Internet Bug Bounty' enables and funds scrutiny of critical internet infrastructure, such as programming languages, web development tools, and communication protocols (refer to Design Lesson 21 for further details).

95 AJL Interview: Marcia Hofmann

96 Such distribution would, of course, require the consent of the people providing reports or research.

97 Notes from conversation between the authors of this paper and Matt Goerzen (September 2021).

As they would not be able to offer researchers the same legal safe harbor assurances as in-house or client-type programs, these independent intermediaries may also need to pursue somewhat innovative solutions to insulate participating researchers and other reporters from legal risk (e.g., by providing access to a communal legal fund contingent on adherence to the intermediary's terms and conditions), while simultaneously advocating for necessary regulatory and legal changes to facilitate greater transparency and more effective redress for harms inflicted by algorithmic systems.

"IF [A] COMPANY IS RUNNING A BUG BOUNTY PROGRAM ON A PLATFORM LIKE HACKERONE AND … YOU MAKE A BEST EFFORT TO FOLLOW THE TERMS, THEN I THINK THAT THERE'S GENERALLY A PRESUMPTION OF GOOD FAITH AND IT … WOULD BE PRETTY DIFFICULT FOR A [LEGA] CASE … TO GO ANYWHERE … [HOWEVER] THERE IS … ALWAYS ROOM TO IMPROVE … FOR INSTANCE, WHERE A BUG BOUNTY POLICY … [HAS] SAFE HARBOR BASED ON … NOT DISCLOSING A REPORT."

— Jack Cable, Security Researcher[98]

Algorithmic systems have the potential to do enormous good for humanity in the coming decades, as well as enormous harm, including through the reinforcement or exacerbation of existing inequities. The kind of institution described above could plausibly help move the needle in the right direction in this domain. However, just as such an intermediary should be one of many ways in which the incentives of algorithmic systems vendors and operators are reshaped, we would note that so too might a more independent platform offering comparable mechanisms in the cybersecurity space help to drive greater accountability in the design, development, deployment, and maintenance of systems on whose security people are often forced to depend. Moreover, while public policy considerations and recommendations largely fall outside the scope of this report, fulsome legal protection for third-party algorithmic harms research and disclosure, including in the context of BBPs and related mechanisms, is ultimately dependent on governments making the right decisions regarding whether and to whom to afford such protection.[99]

"A LOT OF BUG BOUNTY PROGRAMS DON'T NECESSARILY SERVE … TRANSPARENCY AND ACCOUNTABILITY. I'D SAY PROBABLY A MAJORITY OF THEM HAVE SOME KIND OF BLANKET STATEMENT SAYING, 'YOU CAN'T DISCLOSE WITHOUT THE COMPANY'S PERMISSION.' THERE ARE SOME THAT DO A LOT OF DISCLOSURE, … BUT THOSE ARE TYPICALLY … OUTLIERS."

— Jack Cable, Security Researcher[100]

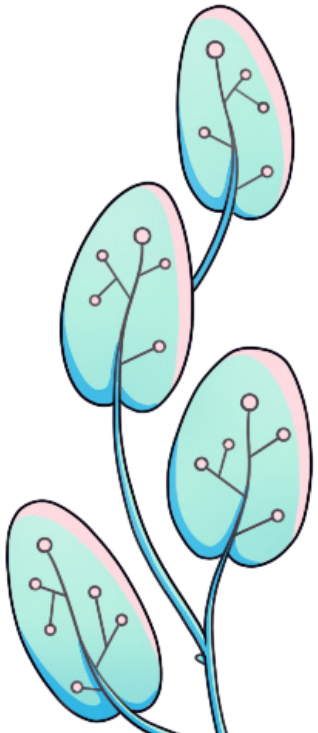
98 AJL Interview: Jack Cable

99 We encourage readers to explore the details of various public policy recommendations from our CRASH project colleagues, including with respect to third-party assessment of algorithmic systems, in "Change From the Outside: Towards Credible Third-Party Audits of AI Systems" (Raji, Costanza-Chock, and Buolamwini Forthcoming).

100 AJL Interview: Jack Cable

"[T]HE PLATFORM MAY HAVE ITS OWN INTEREST TO PROTECT IF A SITUATION REALLY GOES SOUR, AND A COMPANY IS THINKING ABOUT SUING OVER A DISCLOSURE SITUATION. SO I THINK IT WOULD HAVE TO BE REALLY CLEAR THAT THE LEGAL SUPPORT IS [THERE]; PERHAPS THERE'S A FUND THAT CAN BE USED FOR [LEGAL] DEFENSE. BUT I THINK YOU WOULD WANT SOME SEPARATION BETWEEN THE … PLATFORM, AND THE RESEARCHER AND MAKE SURE THAT THE RESEARCHER'S INTERESTS ARE PROTECTED. AND THE PLATFORM'S INTERESTS ARE NOT GUIDING IN THAT SITUATION."

— Marcia Hofmann, Digital Rights Lawyer[101]



101 AJL Interview: Marcia Hofmann

V.
CASE STUDY: TWITTER'S ALGORITHMIC BIAS BOUNTY CHALLENGE

In this section, we explore Twitter's recent "Algorithmic Bias Bounty Challenge" as a case study of both the potential and pitfalls of algorithmic harm BBPs. We briefly survey publicly available background information on the BBP; provide an overview of the program's terms and scoring approach; and apply our design framework to evaluate the program's promises and shortcomings.

BACKGROUND

On July 30, 2021, Twitter's "Machine Learning Ethics, Transparency and Accountability" (META) team shared the exciting news that the company would establish a one-week BBP competition for its image cropping algorithm. This algorithm was used to identify the most 'salient' points in images uploaded by Twitter users and thereby enable automatic cropping of those images around their points of maximum saliency to fit the image frame sizes of Twitter's social feed. The challenge to potential contributors was to identify "what potential harms such an algorithm may introduce … either unintentional … or intentional … affecting anyone from Twitter users to customers or Twitter itself" (HackerOne, 2021b). Three rank-order prizes were offered for the submissions that scored the highest ($3,500, $1,000, and $500, respectively), while two additional prizes of $1,000 apiece were offered to the authors of the "Most Innovative" and "Most Generalizable" submissions.

Twitter was motivated to establish this BBP, according to Chowdhury and Williams' blog post announcing the challenge, in part by its potential value for cultivating a community of researchers and hackers similar to that which exists in the cybersecurity domain. They hoped to "help [Twitter] to identify a broader range of issues than [it] would be able to on [its] own" (Chowdhury and Williams 2021). In addition, the META team pointed to Twitter's prior response to criticism from users regarding this same algorithm, which indicated that it was systematically preferencing crops that included White individuals over Black or Brown individuals, and male-presenting individuals over female-presenting individuals. Three in-house researchers at Twitter had published findings validating this user feedback, and determining that a further line of public critique — that the algorithm would systematically crop pictures of women to emphasize typically objectified physical features, such as their chest or legs, over faces — was not supported by statistical evidence (Yee, Tantipongpipat, and Mishra 2021).

PROGRAM TERMS & SCORING

In detailing how submissions should be structured, and how those submissions would be scored to identify prize-winners, Twitter and HackerOne opted to produce a set of terms and conditions not dissimilar to those established for typical cybersecurity bug bounties, and pointed researchers to the model and cropping code (which was accessible to prospective participants). These program terms also called on contributors to leverage both quantitative and qualitative methods to justify and contextualize identified harms. Twitter developed a first-of-its-kind standardized approach to scoring submissions to its 'bias bounty,' as follows:

1. **Base Point Allocation:** Is the harm intentional or unintentional, and what type of harm is being inflicted?

	Unintentional Harms "Bias, discrimination or related harms that could occur from "natural" images that a well-intentioned user would reasonably post on Twitter."	**Intentional Harms** "Bias, discrimination or related harms that could be elicited from doctored images posted by malicious actors."
Denigration "Situations in which algorithmic systems are actively derogatory or off offensive."	**10 Points**	**20 Points**
Stereotyping "The tendency to assign characteristics to all members of a group based on an over-generalized belief shared by a few."	**10 Points**	**20 Points**
Under-Representation "Under-representation or lack of representation of a sensitive attribute within a dataset category."	**10 Points**	**20 Points**
Mis-Recognition "[T]he action of mistaking a person's identity or failing to recognize someone'[s] humanity."	**7 Points**	**15 Points**
Ex-Nomination "Treating things like whiteness or heterosexuality as central human norms."	**10 Points**	**20 Points**
Erasure "Erasure of representations challenging dominant and harmful narratives of marginalized communities or the erasure of depictions pointing out past harms."	**7 Points**	**15 Points**
Other Harms Reputational (e.g., "damaged public perception"), psychological (e.g., "embarrassment"), economic (e.g., "reduced customers"), etc.	**5 Points**	**8 Points**

2. **Multipliers #1A and #1B – Damage or Impact:** What is the extent of personal harm to members of the population overall, and to marginalized communities?

Average of impact score for (A) marginalized communities and (B) the overall population.

	No disparate impact to marginalized communities	Harm impacting a single dimension of identity / a single marginalized community	Harm impacting multiple dimensions of identity / multiple marginalized communities or the intersection(s) of multiple marginalized identities
1A: Marginalized Communities Multiplier	1.0x	1.2x	1.2x

	Low impact to a person's wellbeing	Moderate impact to person's wellbeing	Severe impact to a person's wellbeing (including illegal / safety-compromising harm)
1B: Overall Population Multiplier	1B: Overall Population Multiplier	1.0x	1.2x

3. **Multiplier #2 – Affected Users:** "[How many] people … are potentially exposed to the [reported] harm?"

	Extremely rare (possible future)	Monthly occurrence (known past and expected future)	Weekly occurrence (known past and expected future)
3: Likelihood Multiplier	1.0x	1.1x	1.2x

4. **Multiplier #3 – Likelihood (Intentional) or Exploitability (Unintentional):** How likely is the harm to occur, or how much "work / skill is required to launch the attack?[102]

	"The attack requires a skilled person and in depth knowledge every time to exploit"	"A skilled programmer could create the attack, and a novice could repeat the steps"	"A novice hacker/programmer could execute the attack in a short time"
3: Exploitability Multiplier	1.0x	1.1x	1.2x

102 With "launch the attack" presumably meaning to "inflict harm."

5. **Multiplier #4 – Justification:** "Is the methodology well motivated? Do authors provide justification for why addressing this harm is important?"

	4: Justification Multiplier
"The methodology is not entirely appropriate for surfacing harms. The authors do not provide context as to why addressing this harm is important or why they approached the problem this way"	0.5x
"The methodology is not well motivated and justification for the significance of the harm is lacking"	0.75x
"The authors provide some justification for why addressing this harm is important. They provide motivation for their methodology"	1.0x
"The authors provide justification for why addressing this harm is important. The methodology is well motivated"	1.25x
"The authors provide strong justification for why addressing this harm is important. The methodology is well motivated and highly appropriate for the task"	1.5x

6. **Multiplier #5 – Clarity:** Does the submission conclusively demonstrate the risk of harm? Are the limitations of the approach properly situated?"

	4: Justification Multiplier
"The methodology is not entirely appropriate for surfacing harms. The authors do not provide context as to why addressing this harm is important or why they approached the problem this way"	0.5x
"The methodology is not well motivated and justification for the significance of the harm is lacking"	0.75x
"The authors provide some justification for why addressing this harm is important. They provide motivation for their methodology"	1.0x
"The authors provide justification for why addressing this harm is important. The methodology is well motivated"	1.25x
"The authors provide strong justification for why addressing this harm is important. The methodology is well motivated and highly appropriate for the task"	1.5x

Leveraging the base score and multipliers, reports were reportedly scored as follows:

$$\textit{Total Score} = \textit{Base Score} \bullet \left(\frac{\textit{Multiplier 1A} + \textit{Multiplier 1B}}{2} \bullet \textit{Multiplier 2} \bullet \textit{Multiplier 3} \bullet \textit{Multiplier 4} \bullet \textit{Multiplier 5} \right)$$

To facilitate the execution of this BBP, the Twitter META team partnered with four judges with expertise in AI and independent security research or hacking — Ariel Herbert-Voss, Matt Mitchell, Peiter "Mudge" Zatko (who is head of security at Twitter), and Patrick Hall — to help adjudicate submissions.

Twitter received 40 unique submissions to its algorithmic BBP from at least 27 contributors (HackerOne, 2021b). The three points-based winners (1st, 2nd, and 3rd places), as well as the "Most Innovative" and "Most Generalizable" award winners were announced on August 8, 2021:

- **1st Place:** Bogdan Kulynych, "whose submission showcased how applying beauty filters could game the algorithm's internal scoring model … [and thereby] amplify … societal expectations of beauty."

- **2nd Place:** Halt AI, "who found the saliency algorithm perpetuated marginalization … [by] reinforcing spatial gaze biases."

- **3rd Place:** Roya Pakzad, who used "bilingual memes … [to show] how the algorithm favors cropping Latin scripts over Arabic scripts."

- **Most Innovative:** Vincenzo di Cicco, who showed that "the algorithm … favored light skin tone emojis … [which] shows how … adjustments to photos can result in shifts to image salience."

- Most Generalizable: An anonymous contributor showed that "adding nearly invisible pixels to an image … [could] alter the algorithm's preferences" (Twitter Engineering 2021).

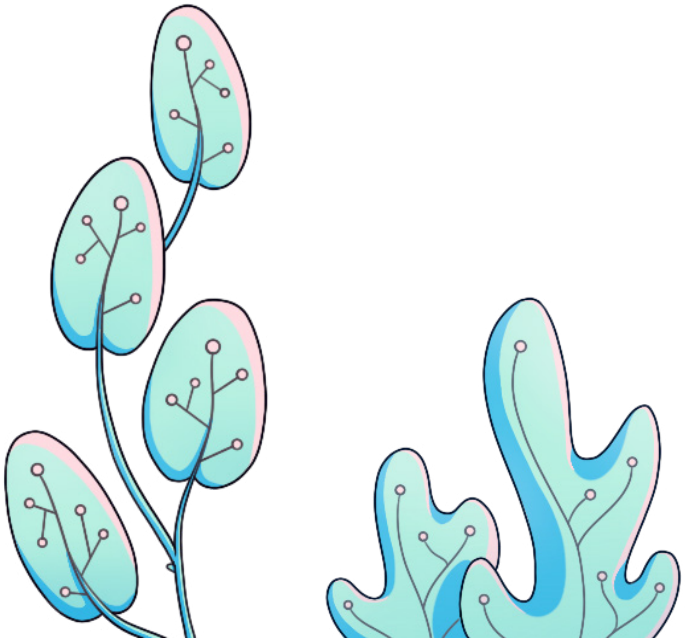
PROMISES AND SHORTCOMINGS OF TWITTER'S FIRST BIAS BOUNTY

Twitter's work in trialing a standardized approach to scoring reports of harmful algorithmic behaviors has potentially wide-reaching value. Myriad organizations may be interested in refining and adopting similar frameworks to handle reports of algorithmic harms surfaced by their users, customers, employees, and others.

Unfortunately, Twitter did not release any of the calculated scores (or sub-scores) for the winning entries, a missed opportunity for transparency around how the scoring framework was applied in practice, leading to limitations on how this framework can be built upon and adapted by others. Ultimately, different organizations may develop their own scoring frameworks. However, the addition of definitions for key terms used in the framework (e.g., affected users, wellbeing, work, skill, attack, exploit, vulnerability, etc.) would better allow Twitter's methodology to be understood, adopted, and evolved by others. Twitter's META team has announced plans to iterate its 'bias bounty' concept, and to adopt elements of the scoring rubric in internal assessments (Yee and Peradejordi, 2021). We hope that in future iterations, key terms and approaches used in the program will be more comprehensively explicated. Similarly, Twitter should look to better distinguish and justify score weightings for harms to institutions, groups, and individuals, respectively. The 'Affected Population Multiplier' in the scoring system rewarded entries for focusing on larger populations, but as Sasha Costanza-Chock noted:

"Thinking about the way that bias and harms in algorithmic systems can play out, it's often — not always, but often — smaller groups of people, maybe people who are multiply marginalized by systems of race, class, gender, disability, and so on, who sometimes can suffer the worst harms from algorithmic systems. As a trans person, I'm from a community that's a very small percentage of the population, and so the scoring mechanism that rewards you for focusing on bigger segments of the population would de-incentivize researchers from focusing on bias problems that might affect trans folks. And…that's just one example" (Field 2021).

In the weeks following the program, Twitter was forthcoming about the bug bounty program having come together quickly, in under a month (Ibid.). This likely did not leave the team much time for engagement with either the algorithmic harms research community, who had already invested in scrutinizing and exposing this system, or with cybersecurity researchers and practitioners, including those, such as Amit Elazari Bar On, who had already been working through how to adapt BBP best practices to other contexts. Relatedly, the program identifies only a small number of resources as having informed its development. Among them, in lieu of adapting CVSS or another widely used risk framework, the "DREAD" (Damage, Reproducibility, Exploitability, Affected Users, and Discoverability) cybersecurity vulnerability scoring framework is referenced (HackerOne, 2021b). This methodology, which was created by Microsoft in the early 2000s, has serious flaws recognized throughout the cybersecurity community. It was "developed without any real academic rigor" (LeBlanc 2007), "add[s] numbers without defining their scales, … making a risk assessment appear algorithmic when it's not" (Shostack 2008, 7), and was "deemed [by Microsoft as] overly subjective and as of 2010 discontinued … in their internal software development lifecycle" (Bodeau, McCollum, and Fox 2018, 21).



The image cropping algorithm seems to have

provided Twitter with an ideal sandbox to test the waters with their inaugural algorithmic bug bounty. We note four pre-existing conditions that likely made this system a 'safe' one to test for the company, providing them with grounds to innovate in how bias bounties can be designed and deployed:

1. **Harms emanating from the image cropping algorithm had already been exposed by users on social media** in fall 2020, almost a year prior to the bounty program being deployed (Ibid.). This state of affairs isn't uncommon; as discussed in multiple places throughout this report, many bounties, including the very first bug bounty by Netscape, have their origin story in users publicly exposing flaws (Ellis and Stevens Forthcoming), and vendors confronting a difficult PR cycle. It also brings a new data point to validate that users do engage in participative experiments with the intent to identify and disclose harmful algorithmic systems, notably on social media.

2. **Twitter had already conducted, and published, a first examination of the ways in which this system was flawed**. In their paper exploring these issues, Twitter researchers Yee, Tantipongpipat and Mishra wrote that after conducting "an extensive analysis using formalized group fairness metrics," they found "systematic disparities in cropping" (Yee, Tantipongpipat, and Mishra, 2021, 1). They identified contributing factors and ultimately advocated for the removal of saliency-based cropping "in favor of a solution that better preserves user agency" (Ibid). The researchers also made the code used for this analysis available to the general public, a meaningful transparency gesture that we hope will become industry norm for similar responses.

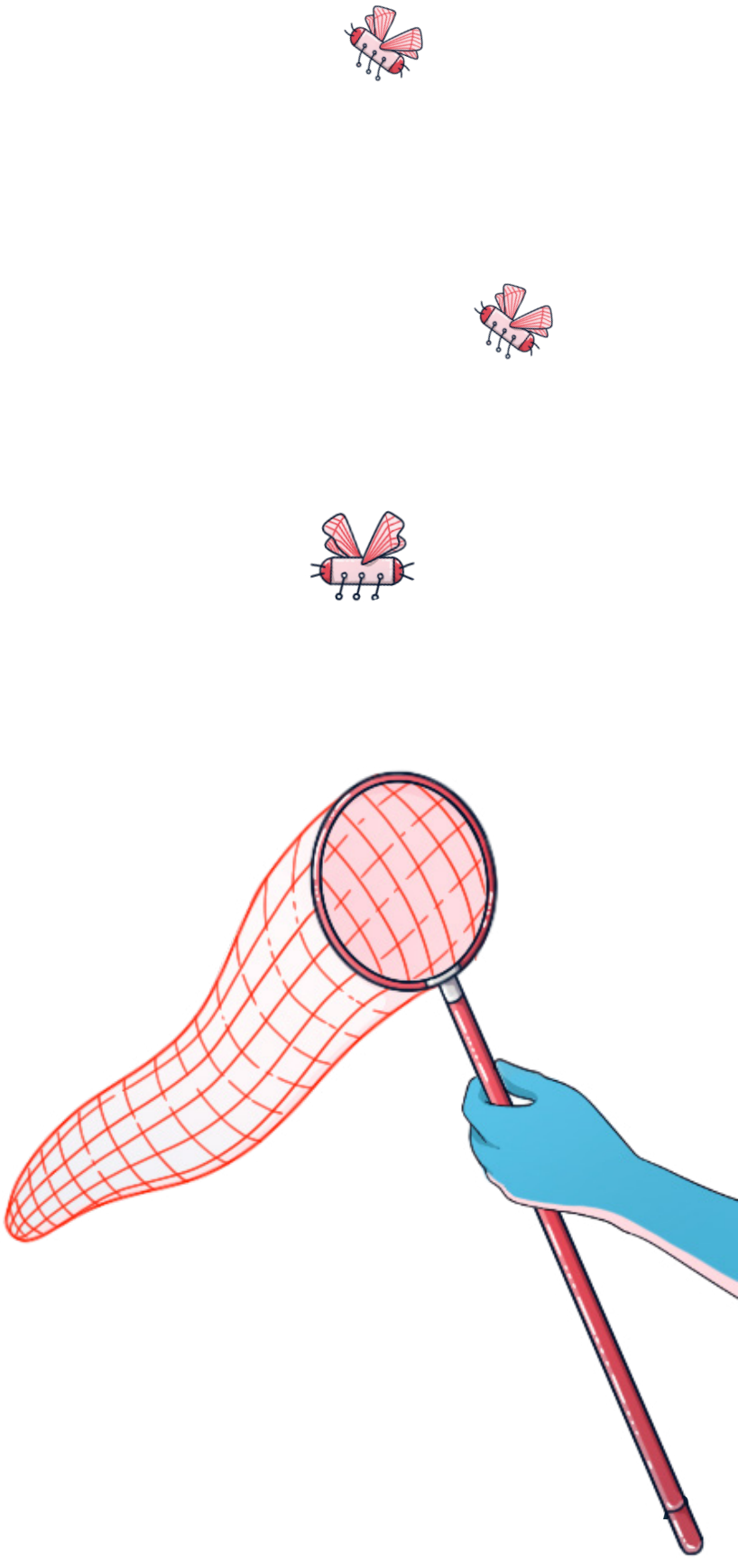
3. As a result of the initial controversy and of the follow-up internal research published, **the company had already taken steps to partially decommission the algorithm**, and officially announced this through its communications department (Field 2021).

4. Finally, **the system was already open source** (Field 2021), meaning that Twitter did not have to make a proprietary system available for research, and could simply draw researchers' attention to an already-open system to drive further examination.

Figure 4: Design Levers Analysis of Twitter's Algorithmic Bias Bounty

TARGET ENTITIES	**Voluntary**		**Adversarial**	
	Only reports relating to organizations that have consented to receiving such reports are accepted.		Reports related to organizations that have not agreed to receive such reports are also accepted.	
COMPENSATION MODEL	**Non-Monetary**	**Bounties**	**Contract**	**Employment**
	Security researchers receive only non-monetary benefits in exchange for their findings.	Organizations or platforms pay security researchers for finding in-scope vulnerabilities, with rediscovery of already-identified vulnerabilities genverally not rewarded.	Organizations or platforms retain security researchers on a temporary, contractual basis to undertake specific services. Researchers are compensated regardless of findings.	Organizations or platforms retain security researchers on a permanent basis to undertake wide-ranging research. Researchers are salaried and receive typical employment benefits.
DISCLOSURE MODEL	**Delayed Full Disclosure**	**Coordinated Disclosure**	**Non-Disclosure**	
	Researchers can freely disclose their findings to the public on a predetermined time frame without additional approval from the affected organization.	Organizations contract with a BBP platform to provide specific services within or related to their BBP, while handling other elements in-house.	Researchers cannot publicly disclose their findings.	
PARTICIPATION MODEL	**Public**		**Private / Invite-Only**	
	All researchers are invited to conduct research and submit reports.		Only pre-authorized researchers are invited to conduct research and submit reports.	
PROGRAM MANAGEMENT	**Platform-Managed**	**Mixed Management**	**Self-Managed**	
	Organizations contract with a BBP platform to deliver their BBP, with reports typically channeled through a specific portal on the platform. The platform also provides related services (e.g., report validation, triage, patch verification, etc.).	Organizations contract with a BBP platform to provide specific services within or related to their BBP, while handling other elements in-house.	Organizations handle delivery of their BBP in-house. They accept reports directly through their websites or via a dedicated email address and handle validation, triage, andpatch verification internally .	
PROGRAM DURATION	**Ongoing** ⟷		**Time-Limited**	
	Reports are accepted on a continuous basis for an evolving range of assets.		Reports are accepted for a specified range of assets over a limited time period.	
PROGRAM SCOPE & ACCESS	**Constrained** ⟷		**Expansive**	
	Only a limited variety of vulnerability types and / or systems are identified as in-scope of the program.		All possible vulnerability types and systems are in-scope of the program.	
	'Closed Box'		**'Open Box'**	
	Testing is limited to publicly available resources and tooling, without additional organizational or technical access (e.g., to documentation or source code).		Additional access, either organizational or technical, is provided to enable a deeper level of testing.	

Considering Twitter's 'bias bounty' through the lens of the design levers established in Bug Bounty Programs 101 helps to distill a number of its core dynamics. Twitter was able to capitalize on the short-term surge of interest that tends to follow the unveiling of a new BBP, while minimizing the longer-term costs and challenges that would accompany a program with ongoing coverage of a wider variety of systems. By leveraging HackerOne's reporting platform, they were able to benefit from access to existing infrastructure and a community of technical researchers, while retaining full flexibility and discretion over scoring and rewards. The key points of divergence between the Twitter program and typical BBP structures — ensuring timely full disclosure and offering only a small number of relatively low-ball rewards — can be explained by the oddities of this program as described above. In particular, the program was unusual in that the algorithm in question was known to be deeply flawed, was already being retired, and was non-proprietary. It will be genuinely interesting to observe whether and how the company continues to develop its algorithmic bias bounties program. We are especially eager to learn how they would address questions around contributor compensation, and the trade-off between the external value to be derived from independent scrutiny of revenue-driving, in-production, proprietary algorithms versus the reputational and business risks posed. As discussed in the closing paragraphs of Key Takeaway #5, there may be a role for an independent intermediary in facilitating the success of such an approach.



VI.
CONCLUSION

Despite growing public concern and increased regulatory efforts focused on the harmful impacts of algorithmic systems, researchers, advocates and practitioners focused on discovering and redressing algorithmic bias and harms continue to face a challenging environment. Done well, BBPs and similar mechanisms might support the development of an inclusive community of practitioners, and they may be useful tools to help ensure robust, transparent processes for algorithmic harms discovery and disclosure.

However, ongoing challenges in cybersecurity vulnerability reporting — such as non-disclosure of identified issues; legal risks to contributors despite the adoption of 'safe harbor' clauses; platform capture by vendors and operators; and labor issues — could easily be reproduced, if we are not careful, as we attempt to stand up systems for the effective discovery, reporting, and redress for algorithmic harms.

As we have explored, the BBP concept is itself highly differentiated in practice. There is a great deal more work to be done to fully understand the history of the vulnerability disclosure ecosystem on its own merits, and to assess its value as a model to be partially emulated in other contexts. Vulnerability disclosure mechanisms, including BBPs, offer many constructive and optimistic lessons for the creation of parallel mechanisms in the algorithmic harms space; transferable best practices already exist. Program terms and conditions can be drafted to maximize accessibility. Vulnerability reporting and disclosure platforms have had some success in cultivating and sustaining community engagement, curating information and resources for researchers, and disseminating best practices for system vendors and operators. Institutions and individuals that seek to organize participatory mechanisms for research and reporting on algorithmic harms should draw from all of these success stories. To that end, we have provided a bulleted list of key recommendations in the Executive Summary at the beginning of this report.

We have also created a Design Companion that follows the main body of the report, where we pull out 25 design lessons from BBPs. We believe that discovery and disclosure mechanisms for algorithmic harms should be designed to avoid reproducing the deficiencies of vulnerability disclosure mechanisms in the cybersecurity space, while also drawing on other parallel domains beyond the scope of this paper (e.g., financial auditing, airline and vehicle safety assessments, and more). In particular, if a goal of greater transparency — much needed in the algorithmic harms context — is to be met, and vendors and operators meaningfully held to account, then mechanisms to facilitate independent scrutiny must be designed so as to be financially independent from system vendors and operators. In so many words, they must avoid the problem of corporate capture. Effective mechanisms will also need to address the complex and varied legal risks — both known and potential — faced by researchers and others who seek to investigate and disclose algorithmic harms. Again, this will be more readily achieved if platforms stand apart from the companies that develop and run the systems to be scrutinized. Without some degree of guaranteed public disclosure, there can be no assurance of accountability.

It is not yet clear how algorithmic system vendors and operators will come to perceive the risk of harm that their systems pose, in particular since regulation of this space remains nascent. It may be the case that institutional incentives will eventually line up in favor of reducing the risks posed to a greater

extent than they have to date in cybersecurity, such that prioritizing the avoidance of algorithmic harms will become mainstream. Indeed, with the rising prominence of social justice movements, emerging regulation for algorithmic systems, and the responsiveness of some technology companies to concerns around the harmful impacts of algorithmic systems, there are legitimate reasons for optimism. However, even as we hope for the best, we need to organize and prepare for the sustained development and deployment of algorithmic systems that pose both broad-based and disparate risk of harm. We hope that this project provides some useful guidance to all of those who are interested in building a world where cybersecurity is improved, and where algorithmic systems and socio-technical systems of all kinds are more equitable and accountable.

DESIGN COMPANION: LESSONS FROM BBPS

Throughout the course of our research, we identified roughly two dozen design lessons related to vulnerability reporting and disclosure mechanisms, their institutional and human participants, and the platforms through which they engage. We believe that each of these lessons has value for the development of comparable mechanisms in the algorithmic harms space. We grouped these lessons, which are otherwise presented here in no particular order, into five main themes:

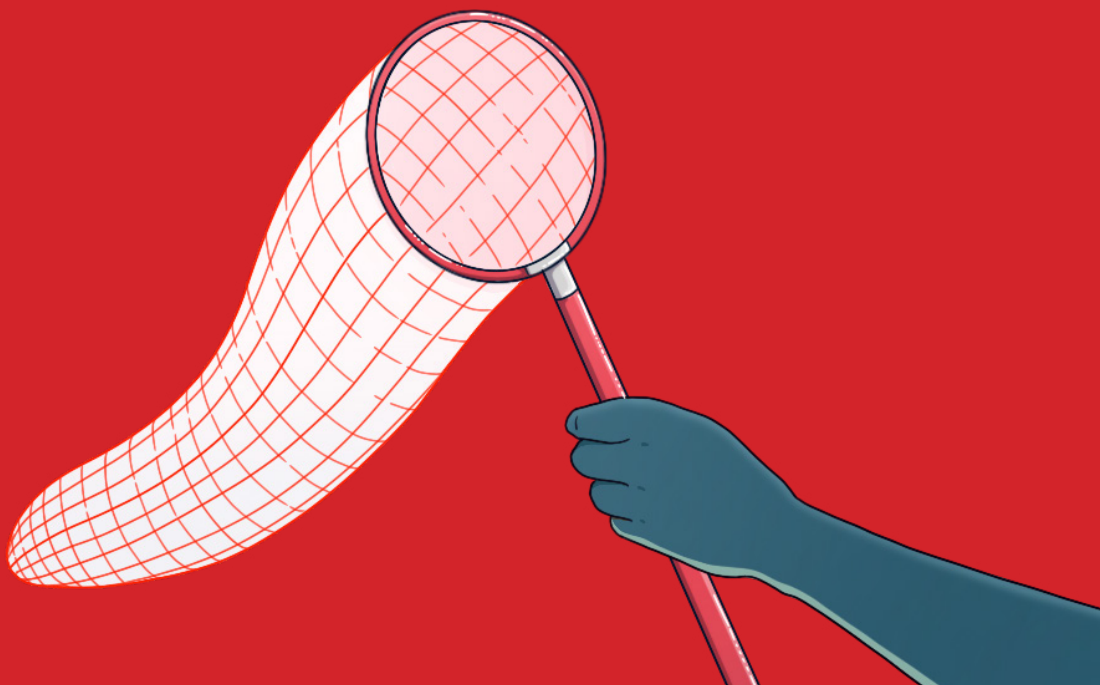
ADVANTAGES OF INTERMEDIARIES

ATTRACTING PARTICIPANTS

ENABLING HIGH-IMPACT REPORTING

ORGANIZATIONAL PREPAREDNESS

FIELD MATURATION



1. RESEARCHERS PARTICIPATE IN BBPS FOR REASONS OTHER THAN THE MONEY

Although BBPs are oftentimes distinguished from similar vulnerability reporting mechanisms by their affordance of monetary compensation for valid reports, financial reward is far from the only motivating factor for security researchers, both in deciding to participate in BBPs and choosing the specific programs they decide to engage with.

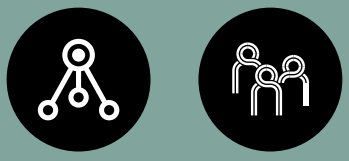
A survey-based study from researchers at the Better Bug Bounties project points to the importance of rewards being both proportionate to research effort and distributed in a timely manner (Akgul et al. 2020, 3-6), as well as other motivating characteristics of BBP programs, such as:

- Wide and diverse program scopes;
- Ease of communication with BBP managers;
- Community perceptions of the BBP; and
- Familiarity, appreciation, or dislike for a product / company.[103]

Overall, it is important to note that only a very small fraction of security researchers make significant income from BBP participation. This widely cited heuristic is backed up by data from HackerOne indicating that fewer than 200 security researchers had earned $100,000 or more in bounties since 2014, when HackerOne was founded (HackerOne 2020a, 4).

While mechanisms for surfacing and reporting algorithmic harms should ensure fair compensation for work undertaken, the monetary value of contributions may also be highly varied.

103 In his interview with AJL, security researcher Dino Dai Zovi provided an experiential affirmation of this survey finding, suggesting that "the people who dig in a little deeper [to discover more significant, complex vulnerabilities] tend to have [an] affinity for what they're looking at … either positive or negative … [if] they think [a company] is gross and they want to beat up on it, or they really like a company or a product."

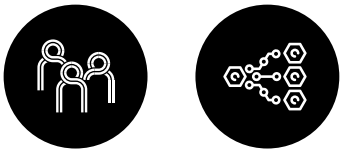


2. REPUTATION OR RANKING SYSTEMS ARE USEFUL FOR BBP PLATFORMS AND RESEARCHERS

Reporting and disclosure platforms make extensive use of reputation-based rewards that are based on researcher-specific characteristics such as number of valid reports submitted, severity of security issues identified, and reliability of report quality / validity.

This reputation data can feed into ranking systems, which are useful for target organizations and platforms to identify talented and reliable security researchers, and for those researchers to compare among their peers and observe their successes reflected in a rising reputation and overall rank, which may afford benefits like access to selective programs or additional financial incentives, such as annual bonuses.

For specific achievements (e.g., submitting reports to 20 different BBPs), participants may also be awarded virtual badges (HackerOne, n.d.) or earn a spot in leaderboard-style halls-of-fame maintained by BBP platforms or individual programs. Such practices could be easily exported to the algorithmic harms context.



3. EVENTS, EDUCATION, AND OTHER INCENTIVES OFFERED BY OR AROUND BBPS CAN HELP COMMUNITIES OF PRACTICE TO GROW

Reporting and disclosure platforms offer a variety of special incentives, events, and educational opportunities designed to accomplish several different goals around building and maintaining participation.

- Engagement — e.g., "Bring a (trusted) friend" policy for HackerOne hackathons;[104] iDefense Referral Program (iDefense 2006).
- Education — e.g., Hacker101 (HackerOne 2018); Bugcrowd University (Bugcrowd, n.d.); YesWeHack Dojo (YesWeHack 2020).
- Identification — e.g., HackerOne "Capture the Flag" activities (HackerOne 2020b); BBP platforms' reputation-based ranking systems
- Retention — e.g., iDefense Retention and Growth Programs (iDefense 2006); invite-only BBPs based on platform ranking

In working collaboratively, for example ahead of hackathons, security researchers may leverage a combination of specialized (e.g., custom hacking tools, Burp Suite, etc.) and generic (e.g., Google Docs / Sheets, Slack, Discord, etc.) software to enable investigation, analysis, and collaboration. Such convenings, as well as the introduction, talent identification, and talent retention practices of reporting and disclosure platforms may map closely onto desirable practices for institutional players in the algorithmic harms space.

These considerations are further discussed in other Design Lessons, as well as Key Research Takeaway #2.

104 AJL Interview: Mårten Mickos

4. BROAD, DIVERSE BBP SCOPES CATER TO THE WIDEST RANGE OF RESEARCHERS

Program terms establish the systems, vulnerability types, and research methods authorized as 'in-scope' of that BBP. A wider, more diverse scope is generally favored by researchers, which requires the organization behind the BBP to have sufficient resources and robust processes for handling the commensurately higher level and variety of reports.

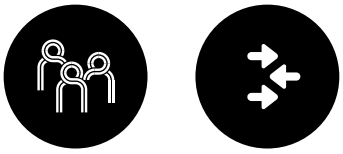
Broad scopes have been theorized to reduce duplication between internal and external security testing (Malladi and Subramanian 2020, 33), while wide and interesting scopes are a key motivating factor for researchers choosing among a range of BBPs (Akgul et al. 2020, 4-6). Such findings could plausibly transfer to the algorithmic harms context.

5. "BUSY WORK" AND OTHER FRICTIONS CAN ACT AS BARRIERS TO PARTICIPATION

Beyond the kinds of 'hard' constraints on participation that BBPs may put in place (e.g., based on reputation, geographic location, background check, etc.), there are a number of 'softer' hurdles that, intentionally or not, may discourage participation. These kinds of frictions include requirements to create test accounts for different BBPs, having to complete CAPTCHA tests, dealing with connection timeouts, and so on, which can be frustrating to would-be contributors (Akgul et al. 2020, 4-6).

Ultimately though, organizations running BBPs need to have a sense of the value of the "long tail" of contributions from occasional participants. In the cybersecurity context, those contributions may ultimately require more resources to manage (notably in triaging reports) than they offer in terms of value to the target organization (Ellis, Huang, Siegel, Moussouris and Houghton 150). This dynamic leads Ellis et al. to conclude that, given the bifurcated labor pool and disproportionate value of the contributions from a small subset of top-tier researchers, the deployment of various constraints on participation, such as invitation-only participation, "may be beneficial and [in the aggregate] carry little downside" (Ibid. 150).

Participatory reporting of algorithmic harms may not follow these dynamics, in the same vein as described in the Design Lesson #14 on invalid and duplicate submissions. Indeed, there may be great value in allowing one-off participants to submit reports, including if these reports are duplicative of issues already documented by others.



6. PEOPLE WANT TO HACK THE ORGANIZATIONS THAT THEY'RE FAMILIAR WITH

Another motivating factor for participation in particular BBPs, is existing knowledge or awareness of particular companies and their products.

In a 2020 survey of successful bug hunters, HackerOne identified liking a company (28% of respondents), being a user or data subject of a particular company (23%), evaluating the technology to buy or use (10%), and even dislike for a company (4%) as factors motivating how security researchers choose their targets (HackerOne 2020a, 39).

As such, BBPs or similar mechanisms can be useful as a means of channeling public awareness (good or bad) into meaningful improvements to the security of widely-used technologies. The same logic applies in the algorithmic harms context.

7. PLATFORMS THAT COORDINATE OR ORGANIZE INFORMATION ABOUT MANY PROGRAMS MAKES IT EASIER FOR RESEARCHERS TO FIND, COMPARE, AND PARTICIPATE IN BBPS

The rapid rise of reporting and disclosure platforms, as well as survey data on what makes them useful for security researchers, indicates that researchers value the role that platforms play in providing access to many programs through a single portal, and allowing contributors to manage their reports and communications with target organizations through a common interface (Akgul et al. 2020, 4-6).

Such portals may also include summary-level program information. For example, YesWeHack's FireBounty aggregator distinguishes VDPs and BBPs, and each program entry includes a summary of the program scope as well as a link to full program terms (YesWeHack, n.d.), while HackerOne's BBP listing provides program-specific statistics (e.g., on reward amounts, response timeliness, etc.). Beyond intermediating platforms, vendors who operate large-scale BBPs are also investing in ensuring relevant information is well-curated and easy to navigate. In the summer of 2021, Google rolled out its "Bug Hunter" dedicated platform (Google 2021), featuring prominent characteristics of platforms such as HackerOne and Bugcrowd, for example a "bug hunter university", a "leaderboard," etc. In essentially replicating these key features, Google presumably aims to help participants navigate among the multiple programs that the company operates.

Especially if a large number and / or wide variety of reporting mechanisms emerge for algorithmic harms, as opposed to a model in which reporting is entirely centralized from the participant perspective, this same kind of dashboard could be immensely useful for established participants and newcomers alike.
multiple programs that the company operates.

8. TIMELY AND WELL-DESIGNED COMMUNICATION ARE CRITICAL TO A STREAMLINED AND PRODUCTIVE REPORTING EXPERIENCE

For organizations to make the reporting of vulnerabilities as frictionless as possible, clearly identifiable, accessible, and trustworthy means of back-and-forth communication with security researchers are a necessity. As Lisa Wiswell Coe, Fmr. Program Manager for Hack the Pentagon noted, "The trick is ... to make sure that you're communicating with [researchers] effectively and in a really professional but kind way."[105]

A degree of standardization across programs in how such communication occurs is also desirable (e.g., use of 'security@[organization].com' as a standard point of contact), so that researchers don't have to navigate an entirely new reporting experience for every program in which they participate. The more numerous the points and means of contact available (e.g., email, social media, web portal, messaging apps, etc.), the greater the logistical challenge for both the researcher and the relevant organization in ensuring that reports end up in the right hands. However, such considerations require careful balancing with the accessibility of reporting channels to different potential participants (e.g., in geographic locations where internet access may be limited).

Thoughtful consideration of what technical infrastructure is most well-suited to secure and verifiable communication is also important (e.g., use of encrypted communication channels for disclosure). In the context of algorithmic harms this is most relevant to protecting contributors from retribution and preserving the confidentiality interests of those who have experienced harm, whereas in the cybersecurity context it is most relevant to preventing third parties from accessing sensitive information related to systems or vulnerabilities.

Overall, organizations and BBP platforms need to put in place appropriate resources and internal processes for responding to reports and researcher questions. Regardless of the specific mechanisms employed, it is crucial that organizations work to make this communication with researchers as easy as possible, since slow or frustrating communication has been recognized as discouraging research and reporting (Akgul et al. 2020, 4-6).

While some organizations may opt to communicate manually, others, such as GitLab, make use of automated response tools to provide initial estimates of how long a researcher should expect to have to wait for their report to be validated, triaged, and acted upon (Strike 2019).

105 AJL Interview: Lisa Wiswell Coe

9. NEEDED: EFFECTIVE DIGESTIVE SYSTEMS AND A STRONG STOMACH

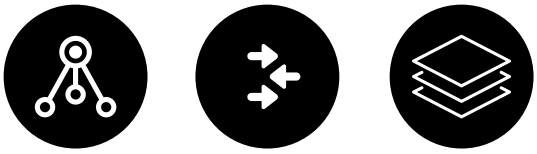
Significant prior investments throughout the product lifecycle — in development, in maintenance, and in iteration and learning — are needed to maximize the value of reporting through BBP-like mechanisms. Without such investment, internal response teams are likely to be inundated with technically actionable findings that they lack the resources and processes to remediate, while key lessons identifiable from such reports that could improve future product development are unlikely to be identified or adopted.

As has been well-documented with respect to cybersecurity, an algorithmic BBP (or comparable mechanism) should not be an organization's first undertaking on a pathway towards more equitable and ethical algorithmic products or services. Rather, systems should be thoughtfully developed and tested prior to deployment to preempt potential harms, which in turn requires significant prior investment in people and processes.

If easy-to-find, easy-to-avoid vulnerabilities are not dealt with ahead of BBP rollout, then those kinds of vulnerabilities tend to make up the overwhelming majority of report submissions, causing administrative overhead and program costs to spike, and leaving the organization's security professionals "thinking that a job as an elephant vasectomist sounds … better" (Moussouris 2018). While duplicate or repeat reports may provide a greater level of benefit relative to cost in the algorithmic harms space, for these kinds of mechanisms to function efficiently at scale, the lowest-hanging fruit has to have been cleared out before participation is scaled up.

In addition, target organizations need resources and processes for responding to reports, conducting root cause analysis for identified issues, and integrating lessons learned into feedback loops to improve future development (refer to Key Takeaway #2). Platforms can enable target organizations to shift some program management functions upstream (e.g., report validation), however, target organizations are uniquely positioned to actually address report findings and drive improvements based on identified problems.

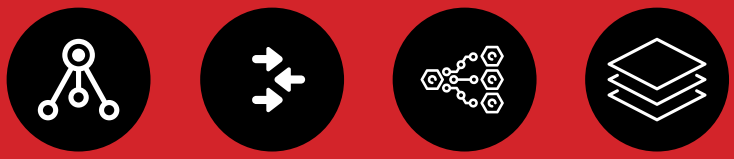
Financially, as well as in many cases reputationally, system vendors and operators need to be willing to take on short-term costs associated with maturing and resourcing their product life cycles to support equity and accountability of algorithmic systems, just as they have to improve security, if they are to realize longer-term benefits such as regulatory readiness, enhanced public or customer trust, and inclusivity-driven innovation. Such a tradeoff takes leadership, and a strong stomach, to commit to and stick with.



10. PUBLISHED VULNERABILITY REPORTS AND OTHER COMMUNITY RESOURCES CAN HELP TO CULTIVATE KNOWLEDGE AMONG PARTICIPANTS AND DRIVE FUTURE IMPROVEMENTS TO REPORT QUALITY

Reporting and disclosure platforms' repositories of past vulnerability reports provide a sizable and varied library of models on which future reports for similar vulnerabilities can be based. They also serve as a useful resource for researchers to further their understanding of vulnerabilities and how different kinds of vulnerabilities can be exploited (Zhao 2016, 84-86).

Both bug hunters and BBP staff are active players in the security research social media ecosystem (featuring tutorials, discussions, and so on). Refer to Key Research Takeaway #3 for further details.



11. REPORTING FORMS OR TEMPLATES CAN BE A USEFUL TOOL TO SHAPE THE RESEARCH PROCESS AND NORMS, AND ENSURE THAT REPORTS CONTAIN ALL RELEVANT DETAILS

To put together a vulnerability report that meets the standards and expectations of recipients, researchers need to understand what formatting, content, and level of detail are expected up-front, so as to minimize the risk of wasted effort or delivery of a flawed submission that would require subsequent revision to be accepted as valid. Clearly articulated, such expectations (e.g., for proof-of-concept in the cybersecurity space) can help to shape research priorities and objectives.

The usefulness of providing standardized templates that researchers can leverage to understand these expectations has been understood since at least the early 2000s, as evidenced by iDefense's early reporting templates (iDefense 2002). These templates can include prompt-style questions and may adapt based on the type of security issue being reported (e.g., vulnerability versus API abuse), such that selecting between different issue types at the top of a form brings up only the most relevant fields for documenting and understanding that issue. Including or requiring completion of certain fields, or providing example content for reports within the templates, sets expectations for the research process and, over time, can influence the shape of the research field overall.

Choices made in the early development of standardized templates or forms for algorithmic harms reporting may set precedents that endure long into the future. The needs of both target organizations in responding to reports and of researchers and other stakeholders, especially impacted communities, should be taken into account from the start.

12. STANDARDIZATION CAN BE USEFUL, BUT ALSO REQUIRES MAKING SUBJECTIVE CHOICES THAT WILL ULTIMATELY INFLUENCE RESOURCE ALLOCATION

As referenced in Tackling Cybersecurity Vulnerabilities and Algorithmic Harms, standardized assessment frameworks for cybersecurity vulnerabilities, such as the Common Vulnerability Scoring System (CVSS), are useful for consistent assessment and prioritization of vulnerabilities.

However, critical analysis of CVSS illustrates how such an approach can fail to capture the true risk and variety of real-world harms associated with different vulnerabilities across distinct contexts (Spring et al. 2021). The development and use of that kind of standardized methodology, if not appropriately grounded and complemented by other qualitative and quantitative perspectives, can also induce the misdirection of resources within organizations and field-wide in unhelpful ways.

Growing awareness of these problems over recent years has helped motivate improvements to the status quo of standardization in the cybersecurity domain. For example, novel approaches, such as the Exploit Prediction Scoring System (EPSS), seek to align more closely with real-world likelihood and levels of harm (Romanosky and Jacobs et al., n.d.). Meanwhile, Spring et al. (2021), articulate a series of recommendations for improving CVSS, summarized as follows:

- Frameworks should be rooted in a more complete understanding of how they are or would be used in practice, both in general and relative to their intended purpose.
- Frameworks should be "transparently evaluated and explainable," and when deployed yield consistent results among practitioners for a range of test cases.
- Frameworks should be response-oriented, and sensitive to context (e.g., organizations, people, and risks).

In light of these considerations, it is especially important that the development of standardized scoring frameworks for algorithmic harms is undertaken collaboratively, and with sufficient interdisciplinary engagement to avoid replicating past mistakes in this novel domain.

13. REFERENCE DATABASES OF COMMON, TAXONOMIZED FLAWS HELP TO ENABLE MORE EFFECTIVE MANAGEMENT AND REMEDIATION OF VULNERABILITIES

Once identified by researchers, vendors / operators, or other actors (e.g., governments), each new cybersecurity vulnerability is assigned a unique identifier — a "Common Vulnerability Enumeration" (CVE) number — by one of several approved CVE Numbering Authorities (National Institute of Standards and Technology, n.d.).

After a vulnerability has been confirmed as a new CVE, it is immediately added to the U.S. government-funded "National Vulnerability Database" (NVD). Analysts from the MITRE Corporation, a non-profit which manages a variety of U.S.-government funded research centers and labs, then undertake additional analysis of each CVE (Ibid.), for example:

- Adding relevant references (e.g., to corresponding patches and advisory notices).
- Assigning a CVSS score to the CVE.
- Associating the record with known, affected product(s), as listed in the NVD's "Common Platform Enumeration" (CPE), which covers applications, operating systems, and hardware.
- Categorizing the CVE using the NVD's "Common Weakness Enumeration" (CWE) list of vulnerability types.

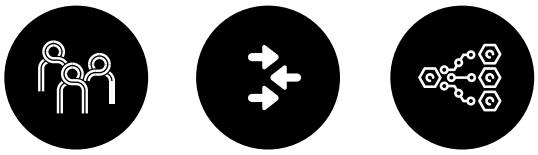
The value of this approach has been recognized for decades in the cybersecurity context; in short, for ensuring a "structured and predictable" method for managing vulnerability findings (Martin, Christey, and Baker 2002). While the precise details of how a comparable mechanism for algorithmic harms should be institutionalized, funded, and maintained would likely diverge from this model, the same kinds of benefits could be expected to accrue from such a systematization and taxonomization of harms reporting. More work is needed to consider how best to structure further, associated details (e.g., particular products, classes of products, or harm types) in a comparable and appropriate manner.

14. INVALID AND DUPLICATIVE REPORTS INCREASE BBPS' ADMINISTRATIVE OVERHEAD, ALTHOUGH NON-NOVEL REPORTS CAN YIELD USEFUL INSIGHTS

Validation of vulnerability reports involves filtering out submissions that do not meet the requirements for a security response or bounty consideration. This requires considerable organizational capacity on the part of the vendor, operator, or other triaging entity and, naturally, demands greater resources as the number of reports to be validated increases. However, these kinds of reports are not without some benefits, which may be even greater in the algorithmic harms context.

Beyond filtering out reports that are spam, inaccurate, or otherwise invalid, the report intake process also needs to be set up to deal with duplicates. In vulnerability research, duplicates may arise during the inevitable gap between when a report is initially submitted and when a fix is publicly issued. Shortening the report-to-fix window is therefore one of the most effective ways of reducing the number and administrative impact of duplicate reports, although the challenge of doing so for complex organizations and systems is significant (Zhao 2016, 102).

Despite the negative impact of duplicate reports, they do have value in illuminating patterns of vulnerability identification and distribution of researcher expertise (Zhao, Laszka, and Grossklags 2017, 381-382). In the algorithmic harms context, reports that document the same harm may in fact be necessary for demonstrating consistently problematic algorithmic decision-making or sustaining public attention on harmful technologies.



15. BBP ROLLOUTS (AND PROGRAM CHANGES) CAN BE TIMED FOR MAXIMUM SECURITY IMPACT

BBPs should integrate with broader efforts to improve product security and should therefore not be understood as fixed entities; rather, program focus and in-scope assets should change over time as organizations' systems and technology offerings evolve. These changes each occur as discrete steps — with the most significant step of all being the creation of a BBP where none existed before.

When a new BBP is created, levels of vulnerability reporting tend to quickly hit their overall peak, before declining over time to a stable rate (Walshe and Simpson 2020, 41). The comparative ease with which bugs can be identified after initial rollout of a BBP also leads a larger number of security researchers to engage with that program (Maillart et al. 2017, 86-87). Substantial, participation-increasing changes to program terms (e.g., to include new systems or vulnerability types) can reasonably be expected to have a similar impact. Because of these dynamics, organizations may wish to target rollout of (or substantive changes to) their programs to align with security testing and bug-fixing during product development. Time-limited BBPs or crowdsourced penetration tests may also be useful as a means of directly targeting security scrutiny towards products or infrastructure at key moments. As security researcher Jack Cable noted: "starting smaller and slowly increasing scope can be helpful for both encouraging participation over time, as well as ensuring that organizations aren't overwhelmed initially with reports."[106]

106 AJL Interview: Jack Cable.

16. BBP PLATFORMS OFFER SERVICES THAT MAKE THE BBP MODEL WORK BETTER FOR MORE ORGANIZATIONS

Reporting and disclosure platforms are not merely channels for reporting security concerns; they can also provide complementary resources and processes — notably, report validation and triage.

These two functions together transmute an unorganized and constantly growing pool of unverified vulnerability reports into an ordered, prioritized, and verified stack of vulnerability reports. The reports that require the most immediate attention are thereby clearly identified, and those that are duplicative or otherwise invalid removed. Historically, this work was undertaken by the target organization itself; however, platforms such as HackerOne, Synack, and Bugcrowd today perform some elements of these functions in advance of passing the reports along to the organization that actually manages the BBP. Similarly, BBP platforms may arrange for researchers to verify security fixes. These kinds of services reduce the workload for final recipients, but researchers have pointed to the fact that they also could lead to burdensome and non-compensated work for hackers (refer to Key Takeaway #4 to understand how and why BBP participants may be vulnerable to exploitation of their labor). Reporting and disclosure platforms also offer resources to guide organizations on how to most effectively set up BBPs to align with the organization's specific security goals.

To the extent that intermediary platforms may be able to support synthesis, validation, and prioritization of algorithmic harms reports in an ethical and productive manner, such organizations may usefully emulate those practices from cybersecurity reporting and disclosure platforms. They may also be able to offer useful resources to organizations to support the just development and deployment of algorithmic systems.

17. PROOFS OF CONCEPT MAKE REPORTS MORE USEFUL AND COMPREHENSIBLE

Vulnerability reports that detail how a discovered flaw could be exploited to compromise the security of the affected system (i.e. proofs of concept) allow target organizations to more accurately gauge potential impacts and award a bounty commensurate to the true severity of vulnerabilities (Strike 2019; Malladi and Subramanian 2020, 35). Conversely, when reports fail to clearly describe these details, there is a greater likelihood that they will be dismissed as invalid or be accepted "only after many rounds of communication between the [researcher] and an organization's security team" (Zhao, Laszka, and Grossklags 2017, 379).

Oftentimes, the easiest way for security researchers to meet this burden of proof is to include a 'proof-of-concept' exploit (described or written in pseudo-code[107]), which allows whoever is responsible for validating the report to more reliably determine the potential impact of the vulnerability.

A similar model could be useful in the algorithmic harms context for describing how certain characteristics of an algorithmic system can lead or are leading to harmful real-world impacts. Much work remains to be done in helping organizations and researchers develop structured and templated ways to assess, report, and taxonomize these harms.

107 Pseudocode is an informal, plain-text description of an algorithm or program's internal logic.

18. DIFFERENT TESTING APPROACHES HAVE DIFFERENT IMPLICATIONS FOR TRUST

Security testing covers a wide range of approaches, with a core distinction between methods that are 'static,' examining the underlying code of software directly, or 'dynamic,' where software or hardware is tested while in operation.

Source code availability has been identified as having a positive impact on the volume of bugs discovered through BBPs and is a prerequisite to static testing (Zhao, Laszka, and Grossklags 2017, 152). However, because source code is often either proprietary or otherwise sensitive, researchers and organizations may be able to more efficiently foster and maintain trust through additional vetting or verification prior to being provided with greater access for testing. Intermediary organizations of the type described in Key Takeaway #5 may be well-positioned to vet, verify, credential, and vouch for researchers to facilitate access, or otherwise obtain relevant code to support more efficient auditing by the community.

This issue is directly relevant in the context of algorithmic harms research, as discussed in From Cybersecurity Vulnerabilities to Algorithmic Harms.

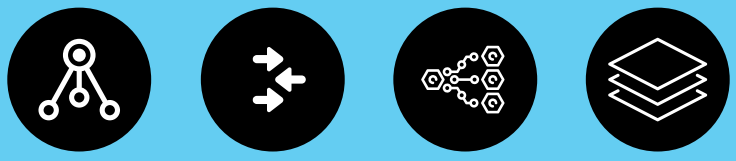
19. BBPS DRIVE SUPPLY AND DEMAND FOR BETTER RESEARCH TOOLING, WHILE MORE MATURE ORGANIZATIONS CAN BENEFIT FROM BBPS FOCUSED ON TOOLS AND ANALYTIC TECHNIQUES

BBPs help to drive supply and demand for greater functionality and availability of security research tools. We can identify this anecdotally in the maturation of security research tools alongside BBPs and related vulnerability reporting models. For example, Burp Suite has gradually evolved since its original 2003 release from a simple tool for manipulating outgoing web traffic into a complex and flexible testing product (Stuttard 2012, 2013; GeeksForGeeks 2019). Security researchers also frequently create their own custom tooling, which can then be shared within the research community (HackerOne 2020a, 31). Refer to Key Takeaway #3 for more details.

Organizations that are already adept at identifying and avoiding common security issues and are therefore looking for ways to efficiently identify more unusual, complex, or novel vulnerabilities should offer bounties to security researchers for sharing custom tools and techniques that they develop (Moussouris and Siegel 2015).

In particular, bounties for 'scripts' that automate repeatable testing processes have begun to emerge over recent years. For example, Mozilla introduced such a component for its BBP in 2019, specifically, bounties for CodeQL queries (Ritter 2019). In determining the bounty eligibility of a submitted script, Mozilla considers its analytic complexity, novelty, and vulnerability identification capability (including instances of false positives), as well as the quality of accompanying documentation and the inclusion of test cases to prove the efficacy of the tool.

This is particularly relevant to the algorithmic harms space, where tools that can be used by communities for participatory audits can be powerful in revealing systematic harms. Rewarding or proactively incentivizing the creation of novel tooling (or other resources) to support algorithmic harms research may be useful for accelerating the development of such resources and maturing the state of the field.

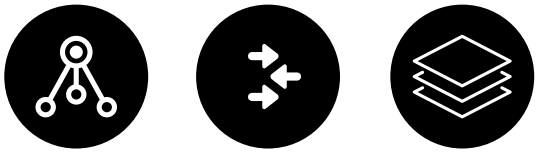


20. DISCLOSE BY DEFAULT

As discussed in Key Takeaway #5, BBPs typically operate under non-disclosure or coordinated disclosure arrangements, through which client vendors ultimately determine whether or not vulnerability information can be publicly disclosed. While the security risks around fixed deadlines for public disclosure are largely addressable through appropriate investment in people and processes, the competitive interests of vendors and operators in exercising control over vulnerability information have impeded the development of 'disclosure-by-default' as a norm.

Nevertheless, the flourishing of the vulnerability disclosure ecosystem more broadly has helped drive best practices around the timing of public disclosure for vulnerabilities. For example, Google's Project Zero has been iteratively evolving its disclosure practices to deliver better cybersecurity outcomes (Willis 2021). BBP-adjacent community initiatives such as Disclose.io (Gallagher 2018) are part of a long legacy of community-led initiatives stretching back into the 1990s (Simple Nomad 1999) that have helped push for consistent disclosure terms across vendors and operators. In a best case scenario, a clear, predetermined disclosure timeline plus regularized coordination between vendors or operators and contributors would reflect an alignment of incentives towards issues being regularly identified and promptly remediated.

Given that algorithmic harms present a different set of exploitation risks (in particular, given that no equivalent offensive market exists for bugs), the risks to wider society of uncoordinated disclosures of algorithmic harms reports are minimal, and the moral case for their public disclosure generally unquestionable. As such, vendors and operators should proactively undertake whatever preparations are necessary for responding constructively to such reports. Aspiring intermediaries in this space should also take heed of similar potential pitfalls, and ensure sufficient resourcing and process maturity to avoid their own bottlenecks (e.g., in achieving redress, triaging reports, responding to participants, etc.).



21. THERE IS A COMMON INTEREST IN REWARDING REPORTS ON 'CRITICAL' TECH INFRASTRUCTURE

A wide range of stakeholders, including governments, corporations, civil society, and the general public, all have an interest in ensuring the security of critical internet infrastructure (CII), such as programming languages (e.g., Python, Ruby), web development tools (e.g., Phabricator, Django), and communication protocols (e.g., OpenSSL).

Because of the widespread interest in the security of this CII, various stakeholders' interests have aligned to support the upkeep of a BBP encompassing major CII elements — the Internet Bug Bounty (IBB) — with the goal of improving "collective safety" by enabling independent security scrutiny of CII (Internet Bug Bounty, n.d.). The IBB operates as a standalone non-profit organization governed by an independent panel of community experts that makes determinations around bounty eligibility and amounts, program rules, and so on. IBB BBPs are administered through HackerOne, while bounties are funded by Facebook, the Ford Foundation, GitHub, Microsoft, and HackerOne.

A similar model could conceivably be implemented to support the maintenance and improvement of critical algorithmic infrastructure, such as widely-used models, datasets, and testing tools.

22. BAD (CYBERSECURITY) DAYS FOR COMPANIES AND GOVERNMENTS ARE GOOD DAYS TO PUSH A BBP

The adoption and evolution of BBPs is closely tied to high-profile 'bad days' for companies and governments in terms of their cybersecurity.

For example, according to Lisa Wiswell Coe, it was only in the wake of the multi-year breach of confidential personnel records held by the U.S. government's Office of Personnel & Management by Chinese state-affiliated hackers (Fruhlinger 2020) that internal advocates for independent security scrutiny of U.S. government systems were able to generate the "political capital" necessary to push through the creation of a BBP.[108] More recently, Facebook's creation of data abuse BBPs — both for its core platform in April 2018 and for Instagram in August 2019 — occurred in the immediate aftermath of high-profile incidents of data misuse by third-party affiliates (Whittaker 2019).

Similarly, in mid-2021, Twitter's Machine Learning Ethics, Transparency and Accountability (META) team conducted its own "bias bounty" as a response to a controversy that emerged on its platform regarding an image cropping algorithm, which had attracted scrutiny previously for systematically preferencing paler-skinned and male-presenting individuals over darker-skinned and female-presenting individuals (refer to Case Study: Twitter's Algorithmic Bias Bounty Challenge).

Around the same time, mobile emulation company Correlium announced its "Open Security Initiative," which plans to select and fund up to three proposals "for research projects designed to validate any security and privacy claims for any mobile software vendor, whether in the operating system or third-party applications" (Corrrelium, 2021). In introducing the initiative, Correlium explicitly pointed to the promises of privacy and security assurance provided by Apple in its plan to analyze images uploaded by users to the company's cloud service, iCloud, in order to detect child sexu (Chowdhury and Williams 2021) al abuse material, which had been swiftly criticized by researchers just days after initial details of the system were unveiled to the public (Mayer and Kulshrestha 2021).

In the cybersecurity context, it is important that an organization's own response to a 'bad day,'

108 AJL Interview: Lisa Wiswell Coe

goes beyond just adopting or pointing to the prior existence of a BBP, which can otherwise be an easy way of providing PR cover in such circumstances without substantially improving the underlying causes of the incident. When accompanied by more fundamental changes in the organization's approach to security, BBPs — especially those that authorize public disclosure on a predetermined timeline — can meaningfully increase transparency and accountability moving forward. Pressure to adopt such measures can come both from within and beyond the organizations themselves, including through the creation of accountability-oriented efforts in the vein of Corellium's Open Security Initiative.

These lessons around timing, and what to push for when an organization's failures are exposed, may be similarly useful for those organizations and individuals seeking to advance the cause of justice in the development and use of algorithmic systems.

23. CLEAR AND COMPREHENSIVE PROGRAM TERMS INCREASE THE VOLUME OF RESEARCH AND REPORTING

Beyond the elements of program terms discussed elsewhere in this section (around BBP scope, communication, legal safe harbor, etc.), both grounded theory-based (Malladi & Subramanian, 36) and quantitative (Laszka, Zhao, and Grossklags 2018, 152) research has demonstrated the importance of clear and comprehensive program terms to researchers' level of engagement with individual programs.

Multilingual program terms may also be useful for bringing in researchers from a wide variety of places and communities. Relatedly, avoiding excessive "legalese" is useful for helping researchers more easily intuit the limits of what research activities are permitted (Malladi & Subramanian, 36-37). These desirable characteristics should be replicated directly in any effort to enable participatory research and reporting on algorithmic harms.

24. RESEARCHERS CHANGE THEIR BEHAVIOR IN RESPONSE TO BOTH REAL AND PERCEIVED LEGAL RISK, WHICH BBP TERMS SHOULD AIM TO MINIMIZE

The legal landscape for security research is generally messy and inconsistent. As digital rights lawyer Marcia Hofmann notes, while much attention is paid to Section 1201 of the Digital Millennium Copyright Act (DMCA) and Section 1030 of the Computer Fraud and Abuse Act (CFAA), "there tends to be a hyper focus on [these laws] in the security research context, but … [it's] important to acknowledge that other laws can be used to intimidate researchers, too; DMCA [and] CFAA aren't the beginning and end of the story."[109] The challenge for researchers of navigating that complexity leads to 'chilling effects' where investigation, reporting, and disclosure of security issues are discouraged. Survey-based research also supports the notion that legal risks, "both perception and reality," are significant enough to alter researchers' behavior and reporting approach (Gamero-Garrido et al. 2017, 1512).

To help researchers navigate legal gray areas and mitigate the associated chilling effect, program terms typically clarify the nature of security research that is permitted. If researchers adhere to these terms, then organizations commit not to pursue legal action against them under the aforementioned laws, or any others. However, while such 'legal safe harbor' clauses provide a degree of assurance to researchers, they typically also stipulate non-disclosure of vulnerabilities and inherently place limits on research scope (refer to Key Takeaway #5).

The challenge of achieving a similar 'safe harbor' mechanism for algorithmic harms research remains open and significant, particularly given the reluctance of companies to authorize comprehensive testing and the necessity of fulsome access for comprehensive research into algorithmic systems. However, in tandem with the approach identified in the discussion of testing in Design Lesson #18, intermediaries may be able to reduce chilling effects in algorithmic harms research through the provision of communal resources, such as a legal aid fund.

109 AJL Interview: Marcia Hofmann

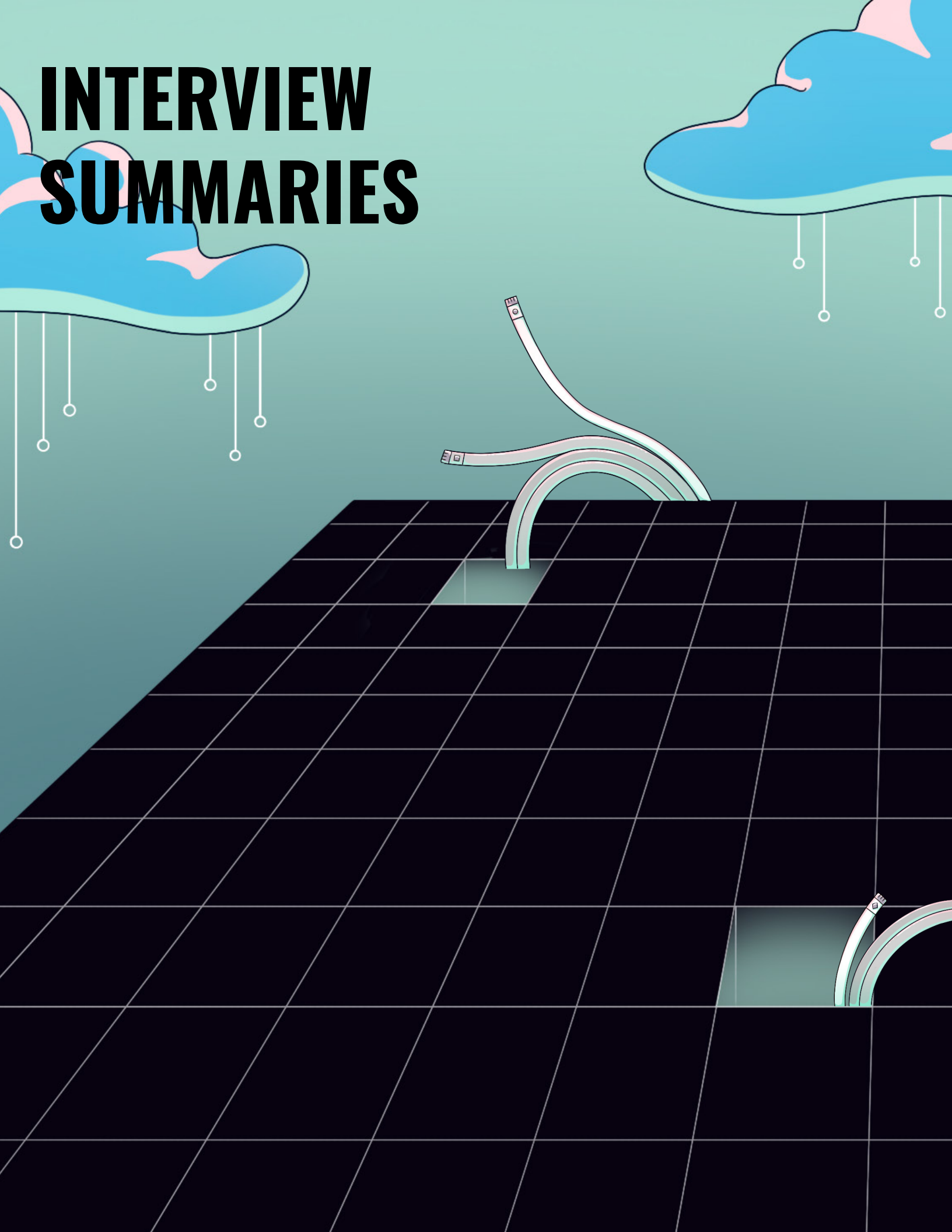
25. PLATFORMS CAN HELP RESOLVE DISPUTES, UP TO A POINT

Reporting and disclosure platforms are, in the abstract, well-positioned to mediate or moderate conflicts between researchers and target organizations (e.g., around bounty eligibility). This is particularly important since target organizations cannot know exactly what issues may be unearthed.

However, to the extent that commercial or reputational interests of vendors and operators shape the incentives and behaviors of intermediaries, the impartiality of those intermediaries and the mediation processes that they put in place is likely to be diminished.

Both researchers and target organizations should have access to mediation processes, and that process should be just and equitable in its resolution of conflicts. Such an approach will be essential for conflict mediation with respect to algorithmic harms reporting.

INTERVIEW SUMMARIES



ALEX RICE

Chief Technology Officer, HackerOne

Image Credit: HackerOne

Alex Rice is a founder and Chief Technology Officer at HackerOne, the world's most trusted hacker-powered security platform. Alex is responsible for developing the HackerOne technology vision, driving engineering efforts, and counseling customers as they build progressive security programs. Alex was previously at Facebook, where he founded and led the product security team. Before that, he was a senior member of the R&D team at Forcepoint, an enterprise security company acquired by Raytheon for $1.9 billion. A long time ago in a state far, far away, Alex worked to improve software & network security as a public servant with the State of Florida.

Rice sees a clear through-line, including from his own personal experiences at Facebook and now HackerOne, from non-compensating VDPs through to BBPs; when converting its VDP to a BBP, Facebook "added two paragraphs" to its the terms and conditions. When Facebook's VDP was first introduced, offering advice to researchers on how to avoid legal risks in reporting, the volume of reports increased 2-3 times. Subsequently, when rewards were added, the volume went up by closer to 20 times. Rice also emphasized that no 'offense-oriented' market exists for the vast majority of software vulnerabilities as they are either too plentiful or can not be usefully exploited, for example, by governments. Therefore, no real competition exists between the majority of BBPs and these 'offense-oriented' markets. The reports generated through VDPs and BBPs are a cost-efficient way of adding feedback loops into product development, which companies spend significant amounts of money generating through other means (e.g., pentesting).

Typically, Rice suggested, too many companies do not voluntarily disclose vulnerabilities by default because they assess the individual risk of each voluntary disclosure to outweigh the derived benefit. The angles taken in some press coverage, in Rice's view, have played a role in sustaining a norm of non-disclosure through inaccurate and / or hyperbolic reporting on individual vulnerabilities. Large organizations are patching multiple unexploited, "near miss" vulnerabilities every week, yet each voluntary

disclosure stands to damage the public's trust in that organization driven by a negative press cycle. Moreover, if any company commits to publishing all reports then Rice suggests that it is likely to find itself at a disadvantage relative to its competitors due to additional scrutiny that can then be applied by prospective customers, partners, or policymakers. Rice believes that normalizing the practice of public vulnerability disclosure would provide a meaningful advantage to cybersecurity defenders everywhere. Public shaming alone is unlikely to change industry norms here, we need collective industry action or government preemption of various risks through legislation or regulation.

Regarding the expansion of BBPs into more socio-technical cybersecurity vulnerabilities or algorithmic harms, Rice pointed out that key impediments are largely internal to the organizations that would need to respond to such reports. BBPs are typically set up and managed by security teams, and those teams may not have the correct competencies or authority to triage and address other types of reports. On the researcher side, Rice affirmed that HackerOne is largely in the business of recruiting and developing traditional security research talent, but that bounties for more socio-technical concerns are already, to a limited extent, occurring organically within traditional, existing BBPs (i.e. without a distinct structure for surfacing these issues). In addition, he indicated that there may be a space for intermediaries to separately and explicitly facilitate the engagement of algorithmic harms researchers in the future.

Note: Rice's interview took place prior to the announcement of Twitter's 'Bias Bounty,' for which it engaged HackerOne to manage submissions.

AMIT ELAZARI BAR ON

Director, Global Cybersecurity Policy, Intel Corporation, and Lecturer, University of California, Berkeley, School of Information

Credit: Amit Elazari Bar On

Dr. Amit Elazari Bar On is a Director for Global Cybersecurity Policy at Intel and a Lecturer at the University of California Berkeley School of Information's Master of Information and Cybersecurity (MICS) program, as well as a member of the External Advisory Committee for the Center of Long Term Cybersecurity. She graduated with a Doctor of Science of the Law (J.S.D.) from UC Berkeley School of Law. During her time at Berkeley she was a CLTC (Center for Long-Term Cybersecurity) Grantee, as well as a member of AFOG, Algorithmic Fairness and Opacity Working Group at Berkeley. Her work on security law, computer crime, privacy and intellectual property has been published in leading law and computer science journals and presented at top conferences such as RSA, Black Hat, USENIX Security, and IEEE Security & Privacy, and featured on leading news sites such as The Wall Street Journal, The Washington Post and the New York Times. Among others', her work was awarded the USENIX Security Distinguished Paper Award, the Annual Privacy Papers for Policymakers (PPPM) Award, Academic Paper Honorable Mention, and the Casper Bowden PET award for Outstanding Research in Privacy Enhancing Technologies. In 2018, she received a Center for Long Term Cybersecurity grant for her work on private ordering regulating information security, exploring safe harbors for security researchers. She practiced law in Israel.

Elazari, unlike some of our other interviews, proposed a different view regarding skepticism of BBPs, especially around labor and legal downsides suggested to be experienced broadly by researchers, as well as the use of NDAs to constrain researchers' ability to freely talk about or publish their findings. While acknowledging these concerns may be relevant in certain cases, and also motivated her work on safe harbor arrangements, she suggested that more quantitative, third-party, broad-scale academic research is needed to assess the breadth and nature of the issues across both these areas of concern. She also pointed to the growth of the BBP ecosystem — including the widespread and growing adoption of legal safe harbor terms and proliferation of security research opportunities to new disciplines and jurisdictions — as being indicative of its relative success in terms of cultivating a community of external researchers. While the BBP economy is still evolving and has shortcomings, she noted that we have seen how it facilitates not just the protection of end users and organizations, but also contributes to skill-building, hiring, diversity, and the development of the discipline and community, while alleviating some of the community's legal concerns associated with vulnerability reporting. We have seen how BBPs helped create a vibrant ecosystem for adversarial hacking and vulnerability research, fueling the creation of a community and even a profession known as "bug hunting".

The concept of an algorithmic bug bounty, as Elazari recalled envisioning it in late 2017 / early 2018, entailed the creation of a market and legal framework for incentivizing scalable, crowd-based auditing through a framework harnessing monetary and legal incentives and drawing from the experience of security research as a tool to empower and scale communities. She noted that such an approach would not be the only path to achieve such scrutiny, but rather would be one 'tool' in the 'tool box'.

Elazari also noted that her prior work, which focused on the development, use, and adoption of standardized contractual terms (with legal safe harbor) as a 'private ordering mechanism' for mitigating some of the potential legal concerns associated with reporting findings under anti-hacking laws, was inspired by the way that Creative Commons has emerged to address comparable copyright issues. She reaffirmed our understanding of the distinct value that can be generated by closed versus open mechanisms, with respect to participation, and compensating versus non-compensating mechanisms. These two types of tools (BBPs and VDPs), serve different functions in an organization's assurance practices, in addition to internal offensive research, for example. They are also not the only tools that can be used to cultivate external research; others include, for example, fostering academic research via academic grants, awards, and collaborations, as well as public policy reforms and multi-stakeholder partnerships.

Through her past research, Elazari has identified several considerations relevant to hypothetical efforts to map the BBP model onto the algorithmic harms space:

- Certain similarity in terms of legal risks (e.g., from the CFAA), which makes the idea of 'legal safe harbor' potentially useful for alleviating some of the potential legal concerns associated with algorithmic harms research (or algorithmic auditing).

- The value of market-like mechanisms as one means of incentivizing research, reporting, participation, and skill development.

- The skill sets needed for algorithmic harms and cybersecurity research are not the same.

- Best practices and industry standards for algorithmic systems are much less mature and robust than in the cybersecurity space, but will need to be developed.

- We may observe organizations exploring combinations of various crowd-sourced frameworks and tools (e.g., open, non-compensating reporting programs similar to VDPs as well "private" algorithmic bug bounties or other closed, compensated research frameworks).

- Having clarity around communications and disclosure expectations is critical.

- Other concepts from cybersecurity (e.g., red teaming) may also be applicable to the discovery and redress of algorithmic harms.

Note: Elazari spoke with AJL in a personal capacity regarding her prior research work.

DINO DAI ZOVI

Security Researcher and co-founder of “No More Free Bugs”

Dino Dai Zovi is a well-known information security industry veteran, speaker, author, and researcher.

Credit: Dino Dai Zovi

Dino is best known in the information security community for winning the first PWN2OWN contest at CanSecWest 2007; presenting his security research at leading conferences such as BlackHat, RSA, and DEFCON; and co-authoring the books “The iOS Hacker’s Handbook,” “The Mac Hacker’s Handbook,” and “The Art of Software Security Testing.” As a long-standing member of the BlackHat community, he is also a member of the BlackHat Review Board as well as a co-founder, organizer, and host of the annual Pwnie Awards.

Looking back on the “No More Free Bugs” initiative, which he co-founded, Dai Zovi acknowledged that while existing mechanisms that would compensate researchers for vulnerability discoveries (e.g., iDefense) did already exist, his goal was for vendors themselves to be sponsoring BBPs, rather than depending upon the goodwill of motivated security researchers to voluntarily report their findings through VDPs. He also noted that the notion of ‘accountability’ for manufacturers was crucial to the early emergence of vulnerability reporting as a practice, given that the only option available for protecting one’s own security as a user was through such accountability.

Dai Zovi also recalled aspects of the history of BBP development that caught him by surprise, notably including the shift from what he had anticipated as a likely focus on desktop software to the predominant focus of today’s BBPs on online software.

Dai Zovi suggested, with respect to translating BBPs for other domains, that a primary consideration should be the availability and utility of a broad ‘cognitive surplus,’ versus narrow, deep, and rare expertise, in those contexts. In addition, he pointed out that sufficient access and proximity to target organizations needs to be provided, relative to the necessity thereof for the research process.

JACK CABLE

Security Researcher

Credit: TJ Stege

Jack Cable is a security researcher and student at Stanford University, a Security Architect at Krebs Stamos Group, and a researcher with the Stanford Empirical Security Research Group. He also works as a hacker at the Defense Digital Service. Before this, Jack served as an Election Security Technical Advisor at CISA, where he led the development and deployment of Crossfeed, a pilot to scan election assets nationwide. Jack joined the Defense Digital Service out of high school, where he helped run the Hack the Pentagon bug bounty portfolio, advised on the next iteration of the DoD Vulnerability Disclosure Program, and built innovative cybersecurity assessment tools. Jack is also a top-ranked bug bounty hunter, having identified over 350 vulnerabilities in companies including Google, Facebook, Uber, Yahoo, and the U.S. Department of Defense. He is ranked within the top 100 hackers all-time on HackerOne.

Cable identified the Bug Bounty Forum, a Slack group which he joined after his first big vulnerability find and now has over 800 members, as the key to his path into the heart of the BBP community. Messaging a HackerOne employee through this group, he was able to get an invitation to an in-person hackathon, where he met future collaborators and first felt himself to be a part of this wider group of hackers. In particular, he noted the transition from being a part of this group in a virtual space to a physical space as being important to his being able to tap into collaborative research with other hackers.

Cable noted that collaboration tends to occur among groups (formal or informal) who have worked together or interacted before. Individuals who have unique skills or understanding of particular companies or products, given the availability of bug reports and generally open-source nature of most hacking tools and knowledge, are a key value-add for collaborative undertakings. Company-specific hackathons now generally involve infrastructure being opened up to participants several weeks in advance of the event itself to allow for more comprehensive research, leading to a greater number of vulnerabilities being discovered and more bounties awarded. Recent developments in hacking tooling have also occurred primarily around this "reconnaissance" work.

Given BBPs' limitations on disclosure, which are often wrapped up with safe harbor protections, hackers are sometimes faced with a dilemma when companies offer both a BBP (by which they can be paid) and a vulnerability disclosure policy (where they aren't paid, but can generally disclose their findings). He would like to see nondisclosure removed from BBP policies more widely.

Cable picked up on a couple of potential areas of shared relevance across both algorithmic harms reporting and cybersecurity BBPs:

- The current state of algorithmic harms disclosure somewhat parallels what the security community has gone through in the past where disclosure would create a major PR debacle and create conflict between researchers and technology companies. Coordinated vulnerability disclosure has been useful for normalizing a system in which companies can acknowledge the existence of vulnerabilities, respond to reports of them, and learn how to avoid them more effectively in the future.

- The idea of legal "safe harbor" may also be relevant, given the ways in which security research and algorithmic harms research may implicate similar laws (e.g., DMCA impediments to reverse engineering).

- Although vulnerabilities might seem like a more concrete concept than algorithmic harms, in actuality assessing the potential impact of a vulnerability is a heavily manual process, which requires researchers and security teams to go through a process of thinking through the security implications of the vulnerability in context.

- VDPs and BBPs are a cost-efficient way of adding feedback loops into product development, which companies spend significant amounts of money generating through other means (e.g., pentesting).

LISA WISWELL COE

Fmr. Program Manager, Hack the Pentagon

Lisa Wiswell Coe is an accomplished leader in the cybersecurity space with a decade of programmatic and cyberware experience. She has served as a strategic advisor to HackerOne and a Principal at GRIMM, a cybersecurity research, engineering, and consulting firm. Previously, Coe worked for the U.S. Defense Digital Service, where she was appointed Special Assistant to the Deputy Assistant Secretary of Defense for Cyber Policy in the Office of the Secretary of Defense and pioneered "Hack the Pentagon," the U.S. federal government's first bug bounty program. She has also served as Technology Portfolio Manager at the Defense Advanced Research Projects Agency (DARPA) overseeing a portfolio of cyber initiatives directly contributing to national security including its flagship cyberwarfare program, Plan X.

Credit: Lisa Wiswell Coe

Coe had extensive experience working with the hacking community prior to leading the launch of U.S. Department of Defense (DoD) BBP. One of the key factors (an "aha" moment, in her words) in government officials' appreciation for the necessity of a BBP was the fallout from the OPM hack, which began in late 2013 and lasted over a year with the compromise of millions of U.S. government personnel records.

However, there was also a lot of work needed to build trust and credibility both internally within the U.S. government and externally, given the legal harassment and derogation that hackers had faced over decades.

This involved making sure that, for hackers, legal protections were watertight, accomplishments were appropriately rewarded, and rules of engagement were clear. The government had to demonstrate credibility

in how it would handle reported information — in good faith and with the intent to actually fix the issues that had surfaced. The DoD also undertook a large-scale PR effort in advance of the first BBP pilot to reinforce their commitment to these goals. Ultimately, it's all about getting the hackers to feel like "part of the team."

Internally, the success of the DoD pilot was tied to its demonstrable impact and, in particular, data on effectiveness and cost-efficiency relative to traditional pen-testing services. These aspects can help drive culture change and increase internal support for BBPs. In a government context, there is also the advantage that once created, programs tend to endure rather than being shut down (unless they are being shut down in order to be expanded). On the one hand, ensuring that contractual language that would prevent organizations from establishing VDPs or BBPs for technology that they operate is avoided, and on the other, including language in other contracts that ensures subcontractors or providers maintain their own VDPs or BBPs. This can be a useful pressure point for driving adoption of BBPs.

MARCIA HOFMANN

Digital Rights Lawyer

Marcia Hofmann is a digital rights attorney, a 2021-2022 US-UK Fulbright Cyber Security Scholar, and a board member at the Filecoin Foundation. In the past, she founded and managed Zeitgeist Law PC, a boutique law firm focused on information security, computer crime, electronic privacy, free expression, and intellectual property issues. Hofmann has also worked at Twitter, the Electronic Frontier Foundation, and the Electronic Privacy Information Center, and has taught courses in computer crime at Colorado Law and internet law at UC Hastings.

Credit: Jennifer Graham

Hofmann believes that establishing clear expectations for both parties in sharing vulnerability information is the best way to avoid conflicts between researchers and target organizations. In addition, she emphasized the role that trusted platforms or individuals can play as mediators and facilitators for these interactions. Engaging through these intermediaries can help to head off potential risks on all sides, for example, concerns about extortion or undue public criticism within target organizations, as well as the threat of non-payment or various forms of legal coercion faced by researchers. However, under the status quo, Hofmann believes that intermediaries' financial dependence on target organizations leads these companies to play an outsized role in defining the terms of engagement for BBPs.

The notion that legal safe harbor language in bug bounty program terms has 'solved' the legal risk to researchers entirely is an oversimplification,

according to Hofmann. She noted, by way of example, that the CFAA is not only a civil law, for which liability is nominally waived under these pseudo-contractual agreements, but also a criminal law, such that the government could potentially bring its own prosecution even if a given company waives a right to pursue legal action. However, the likelihood of such an occurrence, as well as the threat of actual legal action to researchers more generally, are often overstated, in Hofmann's view, and few researchers have ever faced more than a 'cease-and-desist' letter from aggrieved targets. The challenge is therefore to create a sense of practical legal safety, even as we accept that these theoretical risks remain.

For both algorithmic harms and cybersecurity vulnerabilities, there is a need for mechanisms that can generate and maintain trust by allowing researchers to demonstrate that they are operating in good faith, and target organizations to show that they are credible in their commitments to address identified flaws in a timely manner and treat researchers with dignity and respect. To achieve this in the algorithmic harms space, Hofmann believes that thoughtful approaches will be needed to bring large companies on board, adequately protect proprietary algorithms so that they can be comprehensively scrutinized (e.g., with subject-specific NDAs), and ensure that such scrutiny is incentivized for these organizations, for example, through regulatory safe harbors or tax incentives. Intermediaries would need to be similarly creative in attracting, leveraging, and protecting researchers.

MÅRTEN MICKOS

Chief Executive Officer, HackerOne

Mårten Mickos is CEO of HackerOne, the world's leading vendor of hacker-powered security. Mickos is passionate about leadership, distributed teams, global businesses, innovation, infrastructure software, the industry landscape, and technology shifts that have profound impact on society. He believes in openness of tools, methods, and work as this enables the smartest minds of this planet to participate in creating something that is far bigger than the sum of its parts. Mickos was formerly SVP and GM of Hewlett-Packard's Cloud Business, CEO of Eucalyptus, EIR at Benchmark Capital and Index Ventures, SVP and GM of the Database Group of Sun Microsystems, and CEO of MySQL AB.

Credit: HackerOne

Mickos views HackerOne as a platform for elite hacking talent recruitment and development, not as a crowdsourcing platform. Traditional crowdsourcing models rely on workers with homogenous skillsets (e.g., for driving, image tagging), whereas security researchers on HackerOne are highly differentiated in terms of their hacking skill. The skewed distribution of skills leads to a similarly skewed payout model, where a small number of hackers earn the majority of the financial reward. However, both the presence of less skilled, less productive hackers and the noise that they create across report submissions are a natural consequence of the business model. Negative impacts should be managed, rather than seen as a fundamental flaw with the approach itself. Importantly for customers, Mickos noted, HackerOne triages reports on their behalf, meaning that incoming submissions are reviewed, and invalid reports removed. In this way, HackerOne removes "noise" for customers. HackerOne conducts

hackathons, offers training, and runs CTFs that help hackers improve and allow HackerOne to assess their skill development, so they can be invited to more selective BBPs and hackathons in the future.

Having a clearly definable and describable problem space with reproducible issues makes implementation of the BBP platform model more feasible. For example, "Find a bug in X and show how it works" versus distributing via outsourcing the development of software in its entirety. HackerOne's community building emphasizes getting new hackers in the door through multiple channels and giving them resources to develop their skills. When HackerOne hosts in-person hackathons, they sometimes encourage participants to bring a (trusted) friend, thereby outsourcing some of the talent identification function.

Mickos suggested that companies are generally unlikely to undertake one-off bounties, because the decision to start paying for bugs is a bright-line internal policy decision within those organizations. If companies are going to do a BBP, then they need to put sufficient resources into more than just bounty payments in order to go from receiving individual bug reports to improving software development processes overall.

Note: Mickos' interview took place prior to the announcement of Twitter's "Bias Bounty," for which it engaged HackerOne to manage submissions.

Credit: Jennfier Graham

RAYNA STAMBOLIYSKA

Fmr. VP Governance and Public Affairs, YesWeHack

Rayna Stamboliyska focuses on EU cyber diplomacy and resilience including issues related to cybersecurity, strategic autonomy, and data protection. An award-winning author for her most recent book "La face cachée d'Internet" ("The dark side of the Internet", Larousse 2017), Rayna is also an IoT hacker and a staunch proponent of open source, data, and science. At the time of her interview with AJL, Stamboliyska was the VP Governance and Public Affairs at YesWeHack, a global bug bounty and coordinated disclosure leader. Stamboliyska is currently an independent researcher and

freelance writer with a particular interest in social justice & open technologies, as well as the founder of RS Strategy, which advises decision-makers on knowledge technologies, and ICT4D in MENA and the Balkans. Rayna also founded OpenMENA to promote open knowledge in MENA She teaches at Sciences Po Paris and writes up the cybersecurity expert column "50 shades of Internet" at ZDNet.fr. A longtime diversity advocate, she is a Council Member of Women4Cyber Europe.

Stamboliyska emphasized several distinguishing characteristics of French BBP platform YesWeHack including its origins in the hacking and cybersecurity communities, its maintenance of non-revenue generating services (ZeroDisclo and FireBounty) to connect security researchers with target organizations and other disclosure channels, offering special prizes for the quality of report write-ups to incentivize communication skill development, and its use of in-house triage teams as opposed to independent security researchers.

In addition, she noted that so long as systems are imperfect from a security perspective, which is all but guaranteed, the need for mechanisms that support responsibility and accountability will remain — a dynamic that parallels the algorithmic harms space. However, she also pointed to the practical (e.g., tooling, impact scoring) and perceptual challenges (e.g., technical teams or personnel seeing these issues as "bullshit … for the legal people to toy around with") that arise in using these mechanisms to help address more socio-technical harms — specifically in the context of privacy, although again this would map to other domains.

GLOSSARY

Algorithmic harms occur in situations when an actor, such as an individual or an institution, uses an algorithmic system to automate classification, prediction, recommendations, scoring, etc., and as part of a process that causes people harm, such as loss of opportunity, violation of rights, affronts to dignity, social stigma, loss of freedom, physical safety, or loss of life.

Bug bounty programs (BBPs) are mechanisms that incentivize independent security researchers to identify and report 'security bugs' known as vulnerabilities, which exist in software and hardware. Bounties are both reputational (halls of fame, reputation metrics, public credit) and material (cash bounties, merchandise, tickets, coupons).

Cybersecurity vulnerabilities are security flaws that arise as a product of technical deficiencies in digital technologies and/or choices made about how those technologies are deployed and used. Left unresolved, these vulnerabilities can allow for exploitation and hijacking, manipulation, unauthorized access, or corruption of the vulnerable system or network, and ultimately, harm to associated organizations, individuals, or groups.

Penetration testing ("pentesting") is the "business of compromising the security of computer systems at the behest of owners who want to harden their systems from attack" to reveal security flaws that can subsequently be fixed (Perlroth 2021b). Pentesting organizations retain security researchers with specific, demonstrated skills and contract out their labor to clients.

Reporting and disclosure platforms are intermediary organizations that connect security researchers with third-party organizations in order to allow those researchers to disclose relevant vulnerability reports in exchange for some form of compensation — reputational, merchandise, monetary, etc. These platforms vary widely in function and form, but often provide complementary services such as report validation and issue prioritization / triage.

Security researchers are individuals, sometimes computer hackers, who "work, often independently from … institutions, to analyze, explore, and fix … vulnerabilities" (Rodriguez et al. 2018).

Vulnerability disclosure programs (VDPs) are mechanisms that incentivize independent security researchers to identify and report 'security bugs' known as vulnerabilities, which exist in software and hardware. Public credit is generally offered, but no substantial material rewards are provided.

BIBLIOGRAPHY

Aday, Sean. 2018. “Can news coverage of cyber issues get past hacks and attacks?” William and Flora Hewlett Foundation. https://hewlett.org/can-news-coverage-of-cyber-issues-get-past-hacks-and-attacks/.

Akgul, Omer, Taha Eghtesad, Amit Elazari, Omprakash Gnawali, Jens Grossklags, Daniel Votipka, and Aron Laszka. 2020. “The Hackers’ Viewpoint: Exploring Challenges and Benefits of Bug-Bounty Programs.” *6th Workshop on Security Information Workers*, (August). https://www.taahaaa.ir/files/akgul2020hackers.pdf.

Allodi, Luca. 2017. “Economic Factors of Vulnerability Trade and Exploitation.” *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 1483-1499. https://doi.org/10.1145/3133956.3133960.

Altman, Micah, Alexandra Wood, and Effy Vayena. 2018. “A Harm-Reduction Framework for Algorithmic Fairness.” *IEEE Security & Privacy* 16, no. 3 (May / June): 34-45. https://doi.org/10.1109/MSP.2018.2701149.

Anderson, Ross, and Tyler Moore. 2006. “The Economics of Information Security.” *Science* 314, no. 2799 (November): 1-14. http://dx.doi.org/10.1126/science.1130992.

Bacchus, Adam, Sebastian Porst, and Patrick Mutchler. 2019. “Expanding bug bounties on Google Play.” Android Developers Blog (Google Blog). https://android-developers.googleblog.com/2019/08/expanding-bug-bounties-on-google-play.html.

Bodeau, Deborah J., Catherine D. McCollum, and David B. Fox. 2018. “Cyber Threat Modeling: Survey, Assessment, and Representative Framework.” Homeland Security Systems Engineering and Development Institute (HSSEDI) / The MITRE Corporation. https://www.mitre.org/sites/default/files/publications/pr_18-1174-ngci-cyber-threat-modeling.pdf.

Brundage, Miles, Shahar Avin, Haydn Belfield, Gretchen Krueger, Gilian Hadfield, Heidy Khlaaf, Jingying Yang, et al. 2020. “Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims.” arXiv:2004.07213v2, (April), 1-80. https://arxiv.org/pdf/2004.07213v2.pdf.

Brundage, Miles, Sharhar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Defoe, et al. 2018. “The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation.” *arXiv Preprint*, (February), 1-100. https://arxiv.org/pdf/1802.07228.pdf.

Bugcrowd. 2019. *Inside the Mind of a Hacker 2019*. N.p.: Bugcrowd. https://view.highspot.com/viewer/5cdb122666bbaa403f92e406.

Bugcrowd. 2020. *Inside the Mind of a Hacker 2020*. N.p.: Bugcrowd. https://view.highspot.com/viewer/5ef1ad65f7794d0d8ac88cb6.

Bugcrowd. n.d. “Bugcrowd University.” Bugcrowd. Accessed 28, June. https://www.bugcrowd.com/hackers/bugcrowd-university/.

Chin, Monica. 2020. “An ed-tech specialist spoke out about remote testing software — and now he’s being sued.” The Verge, October 22, 2020. https://www.theverge.com/2020/10/22/21526792/proctorio-online-test-proctoring-lawsuit-universities-students-coronavirus.

Chowdhury, Rumman, and Jutta Williams. 2021. “Introducing Twitter’s first algorithmic bias bounty challenge.” Twitter Engineering. https://blog.twitter.com/engineering/en_us/topics/insights/2021/algorithmic-bias-bounty-challenge.

Christen, Markus, Bert Gordijn, and Michele Loi, eds. 2020. “The Ethics of Cybersecurity.” *The International Library of Ethics, Law and Technology / Springer Nature 21*. https://doi.org/10.1007/978-3-030-29053-5.

Cobalt. n.d. “FAQ: Payment.” Cobalt.io. Accessed December 9, 2021. https://cobalt.io/faq/payment.

Coleman, Gabriella. 2012. “Phreaks, Hackers, and Trolls: The Politics of Transgression and Spectacle.” In *The Social Media Reader*, edited by Michael Mandiberg, 99-119. New York, New York: New York University Press. https://gabriellacoleman.org/wp-content/uploads/2012/08/Coleman-Phreaks-Hackers-Trolls.pdf.

Corellium. 2021. “Corellium Open Security Initiative.” Corellium. https://www.corellium.com/blog/open-security-initiative.

Costanza-Chock, Sasha. 2020. *Design Justice: Community-Led Practices to Build the Worlds We Need*. Cambridge, MA: The MIT Press. https://design-justice.pubpub.org.

Craigen, Dan, Nadia Diakun-Thibault, and Randy Purse. 2014. “Defining Cybersecurity.” *Technology Innovation Management Review* 4, no. 10 (October): 13-21. http://timreview.ca/article/835.

Cybersecurity and Infrastructure Security Agency. 2020. “Binding Operational Directive 20-01: Develop and Publish a Vulnerability Disclosure Policy.” Cybersecurity and Infrastructure Security Agency (CISA). https://cyber.dhs.gov/bod/20-01/.

Dai Zovi, Dino. 2009. “No More Free Bugs.” Trail of Bits, Blog (Web Archive). ttps://web.archive.org/web/20090325093503/https://blog.trailofbits.com/2009/03/22/no-more-free-bugs/.

Design Justice Network. 2018. “Design Justice Network Principles.” Design Justice Network. https://designjustice.org/read-the-principles.

Elazari, Amit, Aron Laszka, Omprakash Gnawali, Jen Grossklags, Omer Akgul, Taha Eghtesad, and Daniel Votipka. n.d. “Finding What Motivates Vulnerability Researchers.” Better Bug Bounties. Accessed June 28, 2021. https://vulnstudy.cs.umd.edu/betterbounties/.

Elazari Bar On, Amit. 2018. “We Need Bug Bounties for Bad Algorithms.” *Motherboard (Tech by Vice)*, May 3, 2018. https://www.vice.com/en/article/8xkyj3/we-need-bug-bounties-for-bad-algorithms.

Electronic Frontier Foundation. 2014. “Unintended Consequences: Sixteen Years Under the DMCA.” Electronic Frontier Foundation. https://www.eff.org/wp/unintended-consequences-16-years-under-dmca.

Ellis, Ryan, Keman Huang, Michael Siegel, Katie Moussouris, James Houghton, David L. Shrier, and Alex Pentland. 2017. “Fixing a Hole: The Labor Market for Bugs.” *In New Solutions for Cybersecurity*, edited by Howard Shrobe, 127-160. Cambridge, MA: The MIT Press. https://doi.org/10.7551/mitpress/11636.003.0006.

Ellis, Ryan, and Yuan Stevens. 2021. “Bounty Everything: Hackers and the Making of the Global Bug Marketplace.” *Forthcoming.*

Evans, Mark, Leandros Maglaras, Ying He, and Helge Janicke. 2016. “Human behaviour as an aspect of cybersecurity assurance.” *Security and Communication Networks* 9, no. 17 (November): 4667–4679. https://doi.org/10.1002/sec.1657.

Feathers, Todd. 2021. “Proctorio Is Doubling Down On Lawsuits Against Its Critics.” *Motherboard*, April 29, 2021. https://www.vice.com/en/article/88nae5/proctorio-is-doubling-down-on-lawsuits-against-its-critics.

Field, Hayden. 2020. “An A.I. Training Tool Has Been Passing Its Bias to Algorithms for Almost Two Decades.” *OneZero*, August 18, 2020. https://onezero.medium.com/the-troubling-legacy-of-a-biased-data-set-2967ffdd1035.

Field, Hayden. 2021. “Behind the scenes: How Twitter decided to open up its image-cropping algorithm to the public.” *Morning Brew*, September 27, 2021. https://www.morningbrew.com/emerging-tech/stories/2021/09/27/behind-the-scenes-of-twitter-s-decision-to-open-up-its-image-cropping-algorithm-to-researchers.

FIRST. 2019. “Common Vulnerability Scoring System Version 3.1: Specification Document.” FIRST. https://www.first.org/cvss/specification-document.

Fruhlinger, Josh. 2020. “The OPM hack explained: Bad security practices meet China’s Captain America.” *CSO Online*, February 12, 2020. https://www.csoonline.com/article/3318238/the-opm-hack-explained-bad-security-practices-meet-chinas-captain-america.html.

Gallagher, Sean. 2018. “New open source effort: Legal code to make reporting security bugs safer.” Ars Technica. https://arstechnica.com/information-technology/2018/08/new-open-source-effort-legal-code-to-make-reporting-security-bugs-safer/.

Gamero-Garrido, Alexander, Stefan Savage, Kirill Levchenko, and Alex C. Snoeren. 2017. “Quantifying the Pressure of Legal Risks on Third-party Vulnerability Research.” *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security* (CCS 17), (October), 1501-1513. https://doi.org/10.1145/3133956.3134047.

GeeksForGeeks. 2019. “What is Burp Suite?” GeeksForGeeks. https://www.geeksforgeeks.org/what-is-burp-suite/.

Geiger, Harley. 2018. “Expanded Protections for Security Researchers Under DMCA Sec. 1201.” Rapid7. https://www.rapid7.com/blog/post/2018/11/01/expanded-protections-for-security-researchers-under-dmca-sec-1201/.

Goerzen, Matt, and Gabriella Coleman. Forthcoming. “Many Hats Full Disclosure, Attention Hacking, and the Rise of the Hacker Professional from 1991 to 2000.”

Goerzen, Matt, Elizabeth Anne Watkins, and Gabrielle Lim. 2019. “Entanglements and Exploits: Sociotechnical Security as an Analytic Framework.” *9th USENIX Workshop on Free and Open Communications on the Internet (FOCI 19)*, (August), 1-24. https://www.usenix.org/conference/foci19/presentation/goerzen.

Google. 2021. “Vulnerability Research Grant Rules.” Google Bug Hunters. Accessed October 4, 2021. https://bughunters.google.com/about/rules/5479188746993664.

Google. n.d. “About Project Zero.” Project Zero. Accessed December 9, 2021. https://googleprojectzero.blogspot.com/p/about-project-zero.html.

Greene, Collin. 2018. “Data Abuse Bounty: Facebook Now Rewards for Reports of Data Abuse.” Facebook. https://about.fb.com/news/2018/04/data-abuse-bounty/.

HackerOne. 2018. “Hacker101 – Introduction.” Hacker101. https://www.hacker101.com/sessions/introduction.html.

HackerOne. 2019a. *The Beginners’ Guide to Bug Bounty Program*s. N.p.: HackerOne. https://www.hackerone.com/resources/e-book/the-beginners-guide-to-bug-bounty-programs-1.

HackerOne. 2019b. “Vulnerability Disclosure Guidelines.” HackerOne. https://www.hackerone.com/disclosure-guidelines.

HackerOne. 2019c. “Don’t Believe These 4 Bug Bounty Myths.” HackerOne – Blog. https://www.hackerone.com/blog/dont-believe-these-4-bug-bounty-myths.

HackerOne. 2020a. *The 2020 Hacker Report*. N.p.: HackerOne. https://www.hackerone.com/resources/reporting/the-2020-hacker-report.

HackerOne. 2020b. “CTF – About.” Hacker101. https://ctf.hacker101.com/about.

HackerOne. 2021a. “Bug Bounty vs. Penetration Testing: Differences Explained.” *HackerOne Blog*, June 25, 2021a. https://www.hackerone.com/blog/bug-bounty-vs-penetration-testing-differences-explained.

HackerOne. 2021b. “Twitter Algorithmic Bias.” HackerOne. https://hackerone.com/twitter-algorithmic-bias?type=team.

HackerOne. n.d. “Badges.” HackerOne Platform Documentation. Accessed July 11, 2021. https://docs.hackerone.com/hackers/badges.html.

Hahn, Robert W., and Anne Layne-Farrar. 2006. “The Law and Economics of Software Security.” *AEI-Brookings Joint Center Working Paper 6*, no. 8 (April): 1-63. http://dx.doi.org/10.2139/ssrn.897725.

Hansen, Jacob. 2016. “Deconstructing and Rewiring Bug Bounty Programs.” Cobalt. https://cobalt.io/blog/deconstructing-and-rewiring-bug-bounty-programs.

Hansen, Jacob. 2017. “New Alternatives to Bug Bounty Programs | Cobalt.” Cobalt.io. https://cobalt.io/blog/new-alternatives-to-bug-bounty-programs.

Hupa, Anna, Marc Henson, and Martin Straka. 2021. “Announcing New Abuse Research Grants Program.” Google Security Blog. https://security.googleblog.com/2021/06/announcing-new-abuse-research-grants.html.

iDefense. 2002. “iDEFENSE Security Vulnerability Contributor Program Templates.” iDefense (Archived). http://web.archive.org/web/20021222035846/http://www.idefense.com/contributor_template.html.

iDefense. 2006. “Vulnerability Contributor Program (VCP) Rewards & Incentives.” iDefense (Archived). http://web.archive.org/web/20060211064941/http://labs.idefense.com/vcp_reward_programs.php.

Internet Bug Bounty. n.d. “Internet Bug Bounty.” Internet Bug Bounty. Accessed August 20, 2021. https://internetbugbounty.org/.

Johnson, Khari. 2020. "AI researchers propose 'bias bounties' to put ethics principles into practice." *VentureBeat*, April 17, 2020. https://venturebeat.com/2020/04/17/ai-researchers-propose-bias-bounties-to-put-ethics-principles-into-practice/.

Johnson, Khari. 2021. "What algorithm auditing startups need to succeed." *VentureBeat*, January 30, 2021. https://venturebeat.com/2021/01/30/what-algorithm-auditing-startups-need-to-succeed/.

Jupa, Anna. 2020. "Research Grants to support Google VRP Bug Hunters during COVID-19." Google Security Blog. https://security.googleblog.com/2020/04/research-grants-to-support-google-vrp_20.html.

Laszka, Aron, Mingyi Zhao, and Jens Grossklags. 2018. "The Rules of Engagement for Bug Bounty Programs." Edited by Sarah Meiklejohn and Kazue Sako. *22nd International Conference on Financial Cryptography and Data Security. Lecture Notes in Computer Science (FC 18). Lecture Notes in Computer Science* 10957 (December): 138-159. https://doi.org/10.1007/978-3-662-58387-6_8.

LeBlanc, David. 2007. "DREADful." Microsoft Blog (Archive). https://docs.microsoft.com/en-us/archive/blogs/david_leblanc/dreadful.

Lum, Kristian, and William Isaac. 2016. "To predict and serve?" *Significance* 13 (October): 14-19. https://doi.org/10.1111/j.1740-9713.2016.00960.x.

Mackey, Aaron, and Kurt Opsahl. 2021. "Van Buren is a Victory Against Overbroad Interpretations of the CFAA, and Protects Security Researchers." Electronic Frontier Foundation (EFF). https://www.eff.org/deeplinks/2021/06/van-buren-victory-against-overbroad-interpretations-cfaa-protects-security.

Maillart, Thomas, Mingyi Zhao, Jens Grossklags, and John Chuang. 2017. "Given enough eyeballs, are all bugs shallow? Revisiting Eric Raymond with bug bounty programs." *Journal of Cybersecurity* 3, no. 2 (June): 81-90. https://doi.org/10.1093/cybsec/tyx008.

Malladi, Suresh S., and Hemang C. Subramanian. 2020. "Bug Bounty Programs for Cybersecurity: Practices, Issues, and Recommendations." *IEEE Software* 37 (1): 31-39. https://doi.org/10.1109/MS.2018.2880508.

Marino, Andrew. 2020. "How the commercialization of bug bounties is creating more vulnerabilities," Interview with Katie Moussouris. The Verge. https://www.theverge.com/2020/7/7/21315870/cybersecurity-bug-bounties-commercialization-katie-moussouris-interview-vergecast-podcast.

Martin, Robert, Steven Christey, and David Baker. 2002. "A Progress Report on the CVE Initiative." The MITRE Corporation. https://cve.mitre.org/docs/docs-2002/prog-rpt_06-02/CVE_FIRST_paper.pdf.

Mayer, Jonathan, and Anunay Kulshrestha. 2021. "Opinion: We built a system like Apple's to flag child sexual abuse material — and concluded the tech was dangerous." *The Washington Post* (Washington), August 19, 2021. https://www.washingtonpost.com/opinions/2021/08/19/apple-csam-abuse-encryption-security-privacy-dangerous/.

Metcalf, Jacob, Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh, and Madeline Clare Elish. 2021. "Algorithmic Impact Assessments and Accountability: The Co-construction of Impacts." *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT 21)*, (March), 735-746. https://doi.org/10.1145/3442188.3445935.

Microsoft. 2005. "The Trustworthy Computing Security Development Lifecycle." Microsoft. https://docs.microsoft.com/en-us/previous-versions/ms995349(v=msdn.10).

Miller, Charlie. 2007. *The Legitimate Vulnerability Market: Inside the Secretive World of 0-day Exploit Sales.* N.p.: Independent Security Evaluators. https://econinfosec.org/archive/weis2007/papers/29.pdf.

Moussouris, Katie. 2018. "Crouching Tiger, Sudden Keynote." Dubai, United Arab Emirates: Hack In The Box Security Conference 2018. YouTube. https://www.youtube.com/watch?v=6EvI-DdWJZM.

Moussouris, Katie. 2021b. "SolarWinds and Beyond: Improving the Cybersecurity of Software Supply Chains, Before the Committee on Science, Space, and Technology Subcommittees on Investigations & Oversight and Research & Technology of the U.S. House of Representatives, 107th Congress," (Statement of Katie Moussouris, CEO, Luta Security). https://science.house.gov/imo/media/doc/Moussouris%20Testimony.pdf.

Moussouris, Katie, and Michael Siegel. 2015. "The Wolves of Vuln Street: The 1st System Dynamics Model of the 0day Market." In *RSA Conference 2015.* San Francisco, CA. Presentation. https://ic3-2017.mit.edu/sites/default/files/documents/MichaelSiegelKatieMoussouris_VulnMarketsRSAC2015Speaker.pdf.

Moussouris, Katie (@k8em0). 2021a. "Twitter Post, 3:07 PM Eastern Time, March 14." https://twitter.com/k8em0/status/1371176138362392577.

National Institute of Standards and Technology. n.d. “CVEs and the NVD Process.” National Vulnerability Database. Accessed October 3, 2021. https://nvd.nist.gov/general/cve-process.

Nissenbaum, Helen. 2004. “Hackers and the Contested Ontology of Cyberspace.” *New Media & Society* 6 (2): 195-217. https://doi.org/10.1177/1461444804041445.

Nissenbaum, Helen. 2005. “Where Computer Security Meets National Security.” *Ethics and Information Technology* 7, no. 2 (June): 61-73. https://doi.org/10.1007/s10676-005-4582-3.

Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. “Dissecting racial bias in an algorithm used to manage the health of populations.” Science 366, no. 6464 (October): 447-453. https://doi.org/10.1126/science.aax2342.

Olsen, Mike. 2021. “Mike Olsen to Senators Richard Blumenthal, Chris Van Hollen, Elizabeth Warren, Ron Wyden, Tina Smith, and Cory Booker.” Electronic Privacy Information Center. https://epic.org/privacy/dccppa/online-test-proctoring/Proctorio-senate-response-010721.pdf.

Open Bug Bounty. n.d. “Non-Intrusive, Coordinated and Ethical Testing by the Community and for the Community,” Presentation. Accessed December 9, 2021. https://www.openbugbounty.org/open-bug-bounty/Open%20Bug%20Bounty%20-%20How%20It%20Works.pdf.

Paolillo, Dana. 2017. “Computer Fraud and Abuse Act: Made for International Hackers or Average Internet Users.” *Law School Student Scholarship (Seton Hall University)* 932:1-27. https://scholarship.shu.edu/cgi/viewcontent.cgi?article=1922&context=student_scholarship.

Park, Sunoo, and Kendra Albert. 2020. *A Researcher’s Guide to Some Legal Risks of Security Research.* N.p.: Harvard Cyberlaw Clinic and Electronic Frontier Foundation. https://clinic.cyber.harvard.edu/files/2020/10/Security_Researchers_Guide-2.pdf.

Perlroth, Nicole. 2021a. *This Is How They Tell Me the World Ends: The Cyberweapons Arms Race.* First ed. London, England: Bloomsbury Publishing.

Perlroth, Nicole. 2021b. “Daniel Kaminsky, Internet Security Savior, Dies at 42.” *The New York Times*, April 27, 2021b. https://www.nytimes.com/2021/04/27/technology/daniel-kaminsky-dead.html.

PortSwigger. n.d. “Web Security Academy.” PortSwigger. Accessed June 28, 2021. https://portswigger.net/web-security.

Proctorio. 2021. “Products.” Proctorio. https://proctorio.com/products.

Raji, Inioluwa Deborah, Costanza-Chock Sasha, and Joy Buolamwini. Forthcoming. “Change From the Outside: Towards Credible Third-Party Audits of AI Systems.” In Missing Links in AI Policy. N.p.: UNESCO.

Raji, Inioluwa Deborah, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. “Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing.” Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAccT 20), (January), 33-44. https://doi.org/10.1145/3351095.3372873.

Rash, Wayne. 2019. “Hacker Drops Steam Zero Day After Being Banned From Valve Bug Bounty Program.” *Motherboard* / Vice, August 29, 2019. https://www.vice.com/en/article/wjwd8n/hacker-drops-steam-zero-day-after-being-banned-from-valve-bug-bounty-program.

Rice, Tony, Josh Brown-White, Tania Skinner, Nick Ozmore, Nazira Carlage, Wendy Poland, Eric Heitzman, and Danny Dhillon. 2018. *Fundamental Practices for Secure Software Development: Essential Elements of a Secure Development Lifecycle Program (Third Edition).* N.p.: SAFECode. https://safecode.org/wp-content/uploads/2018/03/SAFECode_Fundamental_Practices_for_Secure_Software_Development_March_2018.pdf.

Ritter, Tom. 2019. “Adding CodeQL and clang to our Bug Bounty Program.” Mozilla Security Blog. https://blog.mozilla.org/security/2019/11/14/adding-codeql-and-clang-to-our-bug-bounty-program/.

Rockstar Games. 2021. “Rockstar Games Bug Bounty Program Policy.” HackerOne. https://hackerone.com/rockstargames?type=team.

Rodriguez, Katitza, Nate Cardozo, Jamie Williams, Ramiro Ugarte, and Tamir Israel. 2018. “Protecting Security Researchers’ Rights in the Americas.” Electronic Frontier Foundation. https://www.eff.org/wp/protecting-security-researchers-rights-americas.

Romanosky, Sasha, Jay Jacobs, Ben Edwards, Idris Adjerid, and Michael Roytman. n.d. “The EPSS Mode.” FIRST. Accessed October 3, 2021. https://www.first.org/epss/model.

Sandvig, Christian, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. “Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms.” *Presented to “Data and Discrimination: Converting Critical Concerns into Productive Inquiry,” a preconference at the 64th Annual Meeting of the International Communication Association*, (May), 1-23. http://www-personal.umich.edu/~csandvig/research/Auditing%20Algorithms%20--%20Sandvig%20--%20ICA%202014%20Data%20and%20Discrimination%20Preconference.pdf.

Satheesan, Akash. 2020. “Keys to the kingdom: Proctortrack.” Proctor Ninja. https://proctor.ninja/keys-to-the-kingdom-proctortrack.

Satheesan, Akash. 2020. “Wave Rake: Proctorio.” Proctor Ninja. https://proctor.ninja/wave-rake-proctorio.

Satheesan, Akash. 2021. “Proctorio’s facial recognition is racist.” Proctor Ninja. https://proctor.ninja/proctorios-facial-recognition-is-racist.

Schneier, Bruce. 2007. “Schneier: Full Disclosure of Security Vulnerabilities a ‘Damned Good Idea.’” Schneier on Security (Reproduced from CSO Online). https://www.schneier.com/essays/archives/2007/01/schneier_full_disclo.html.

Schulte, Stephanie Ricker. 2008. “The WarGames Scenario” Regulating Teenagers and Teenaged Technology (1980-1984).” *Television & New Media* 9 (6): 487-513. https://doi.org/10.1177/1527476408323345.

Shostack, Adam. 2008. “Experiences Threat Modeling at Microsoft.” *Proceedings of the Workshop on Modeling Security (MODSEC @ MoDELS)* 413 (September). http://ceur-ws.org/Vol-413/paper12.pdf.

Simple Nomad, aka. Mark Loveless. 1999. “Announcement.” Nomad Mobile Research Centre (Archive). https://web.archive.org/web/20000816155745/http://www.nmrc.org/advise/policy.txt.

Spring, Jonathan, Eric Hatleback, Allen Householder, Art Manion, and Deana Shick. 2018. “Towards Improving CVSS.” *Software Engineering Institute, Carnegie Mellon University*, (December). https://resources.sei.cmu.edu/asset_files/WhitePaper/2018_019_001_538372.pdf.

Spring, Jonathan, Eric Hatleback, Allen Householder, Art Manion, and Deana Shick. 2021. “Time to Change the CVSS?” *IEEE Security & Privacy* 19, no. 2 (March): 74-78. https://doi.org/10.1109/MSEC.2020.3044475.

Stevens, Yuan, Stephanie Tran, Ryan Atkinson, and Sam Andrey. 2021. “Appendix A: Global State of Play for Coordinated Vulnerability Disclosure.” In *See Something, Say Something June 2021 Coordinating the Disclosure of Security Vulnerabilities in Canada*, 1-50. Toronto, Ontario: Cybersecure Policy Exchange. https://www.cybersecurepolicy.ca/vulnerability-disclosure.

Stewart, Camille. 2020. “Systemic Racism Is a Cybersecurity Threat.” Council on Foreign Relations. https://www.cfr.org/blog/systemic-racism-cybersecurity-threat.

Stewart, Camille, and Lauren Zabierek. n.d. “#ShareTheMicInCyber.” Accessed June 28, 2021. https://sharethemicincyber.splashthat.com/.

Strike, Ethan. 2019. “What we learned by taking our bug bounty program public.” GitLab Blog. https://about.gitlab.com/blog/2019/07/19/what-we-learned-by-taking-our-bug-bounty-program-public/.

Stuttard, Dafydd. 2012. “New Burp Suite Extensibility.” PortSwigger (Blog). https://portswigger.net/blog/new-burp-suite-extensibility.

Stuttard, Dafydd. 2013. “Burp through the ages.” PortSwigger (Blog). https://portswigger.net/blog/burp-through-the-ages.

Supreme Court of British Columbia. 2020. “Proctorio, Inc. v. Linkletter,” Notice of Civil Claim. In the *Supreme Court of British Columbia Vancouver Registry.* https://drive.google.com/file/d/1RoGv1E03h6qagU-mT8QsHVmffaxxAhlO/view.

Synack. 2015. “Disclosure Policy.” Synack. https://www.synack.com/disclosure-policy/.

Synack. n.d. “End User Agreement.” Synack. Accessed June 28, 2021. https://www.synack.com/end-user-agreement/.

Synack. n.d. “Synack Red Team FAQs.” Synack. Accessed June 28, 2021. https://www.synack.com/red-team-faq/.

Tramèr, Florian. 2021. “Does Adversarial Machine Learning Research Matter.” In *KDD 2021 Workshop on Adversarial Machine Learning.* Singapore: Association for Computing Machinery. https://floriantramer.com/docs/videos/advml21.mp4.

Trend Micro. n.d. “Disclosure Policy.” Zero Day Initiative. Accessed December 9, 2021. https://www.zerodayinitiative.com/advisories/disclosure_policy/.

Trend Micro. n.d. “Zero Day Initiative FAQ.” Zero Day Initiative. Accessed December 9, 2021. https://www.zerodayinitiative.com/about/faq/.

Trend Micro. 2021. “Trend Micro’s Zero Day Initiative Enhances Position as World’s Largest Vulnerability Disclosure Player.” Trend Micro Newsroom. https://newsroom.trendmicro.com/2021-05-19-Trend-Micros-Zero-Day-Initiative-Enhances-Position-as-Worlds-Largest-Vulnerability-Disclosure-Player.

Trilling, David. 2016. “Hacking: What journalists need to know. A conversation with Bruce Schneier.” The Journalist’s Resource. https://journalistsresource.org/economics/hacking-bruce-schneier-journalists-cyberattacks-ddos/.

Twitter Engineering. 2021. “Twitter Thread August 9, 2021, 3:48PM ET.” https://twitter.com/TwitterEng/status/1424819778397511680.
U.S. Cyberspace Solarium Commission. 2020. “Final Report.” U.S. Cyberspace Solarium Commission. https://www.solarium.gov/report.

von Solms, Basie, and Rossouw von Solms. 2018. “Cybersecurity and information security – what goes where?” *Information and Computer Security* 26, no. 1 (March): 2-9. https://doi.org/10.1108/ICS-04-2017-0025.

von Solms, Rossouw, and Johan van Niekerk,. 2013. “From information security to cyber security.” *Computers and Security* 38:97-102. https://doi.org/10.1016/j.cose.2013.04.004.

Walshe, Thomas, and Andrew Simpson. 2020. “An Empirical Study of Bug Bounty Programs.” *2020 IEEE 2nd International Workshop on Intelligent Bug Fixing (IBF)*, 35-44. https://doi.org/10.1109/IBF50092.2020.9034828.

Whittaker, Meredith. 2021. “The Steep Cost of Capture.” Interactions 28, no. 6 (November): 50-55. https://doi.org/10.1145/3488666.

Whittaker, Zach. 2019. “After data incidents, Instagram expands its bug bounty.” *TechCrunch*, August 19, 2019. https://techcrunch.com/2019/08/19/instagram-data-abuse-bug-bounty/.

Whittaker, Zack. 2018. “Lawsuits threaten infosec research — just when we need it most.” *ZDNet*, February 19, 2018. https://www.zdnet.com/article/chilling-effect-lawsuits-threaten-security-research-need-it-most/.

Willis, Tim. 2021. “Thursday, April 15, 2021 Policy and Disclosure: 2021 Edition.” Project Zero. https://googleprojectzero.blogspot.com/2021/04/policy-and-disclosure-2021-edition.html.

WordPress Support Forum. 2020-2021. “Thread: OpenBugBounty Security Vulnerability Notification.” WordPress Support Forum. https://wordpress.org/support/topic/openbugbounty-security-vulnerability-notification/.

Yee, Kyra, and Irene F. Peradejordi. 2021. “Sharing learnings from the first algorithmic bias bounty challenge.” Twitter Blog. https://blog.twitter.com/engineering/en_us/topics/insights/2021/learnings-from-the-first-algorithmic-bias-bounty-challenge.

Yee, Kyra, Uthaipon Tantipongpipat, and Shubhanshu Mishra. 2021. “Image Cropping on Twitter: Fairness Metrics, their Limitations, and the Importance of Representation, Design, and Agency.” arXiv, (September). doi.org/10.1145/3479594.

YesWeHack. 2020. “Dojo.” YesWeHack Dojo. https://dojo-yeswehack.com/.

YesWeHack. n.d. “FireBounty.” FireBounty | YesWeHack. Accessed July 15, 2021. https://firebounty.com/.

Zatko, Peter. 2009. “Chapter 1: Psychological Security Traps.” In *Beautiful Security*, edited by Andy Oram and John Viega. Sebastopol, CA: O’Reilly Media, Inc. https://www.oreilly.com/library/view/beautiful-security/9780596801786/ch01.html.

Zhao, Mingyi. 2016. “Discovering and Mitigating Software Vulnerabilities Through Large-Scale Collaboration,” Doctoral Dissertation. The Pennsylvania State University. https://etda.libraries.psu.edu/files/final_submissions/13128.

Zhao, Mingyi, Aron Laszka, and Jens Grossklags. 2017. “Devising Effective Policies for Bug-Bounty Platforms and Security Vulnerability Discovery.” *Journal of Information Policy* 7:372-418. https://doi.org/10.5325/jinfopoli.7.2017.0372.

Zorz, Mirko. 2003. “Interview with Sunil James, Manager of iDEFENSE’s Vulnerability Contributor Program.” *HelpNetSecurity*, April 1, 2003. https://www.helpnetsecurity.com/2003/04/01/interview-with-sunil-james-manager-of-idefenses-vulnerability-contributor-program/.

ALGORITHMIC JUSTICE LEAGUE