

# RSA® Conference 2022

San Francisco & Digital | June 6 – 9

## TRANSFORM

SESSION ID: HUM-R05

# Better Bug Bounties? *Lessons on the Disclosure of Security Vulnerabilities & Algorithmic Harms*

**Josh Kenway**

Tech & InfoSec Policy Program Manager, PayPal  
Fmr. Research Fellow, Algorithmic Justice League  
Twitter: @cybersemantics



# Disclaimer

Presentations are intended for educational purposes only and do not replace independent professional judgment. Statements of fact and opinions expressed are those of the presenters individually and, unless expressly stated to the contrary, are not the opinion or position of RSA Conference LLC or any other co-sponsors. RSA Conference does not endorse or approve, and assumes no responsibility for, the content, accuracy or completeness of the information presented.

Attendees should note that sessions may be audio- or video-recorded and may be published in various media, including print, audio and video formats without further notice. The presentation template and any media capture are subject to copyright protection.

©2022 RSA Conference LLC or its affiliates. The RSA Conference logo and other trademarks are proprietary. All rights reserved.

# Introduction

**The CRASH Project:  
Motivations & Research Overview**



# What is the CRASH Project?

- Community Reporting of Algorithmic System  
Harms = “CRASH”
- Kicked off in mid-2020 by the Algorithmic Justice League (AJL)
- Aims to bring together key stakeholders to inform prototyping of tools for **broader participation** in creating **more accountable, more equitable**, and **less harmful** algorithmic (or “AI”) systems

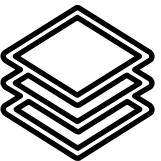


# Why are we (still) talking about bug bounties?

- We set out to explore the viability of creating a **reporting platform for algorithmic harms** similar to ‘bug bounty’ programs and platforms for security vulnerabilities.
- We also focused on how these mechanisms affect and shape:



Community-Building



Field of Practice



Transparency & Accountability

- To better understand these considerations, we explored **design variations** across programs / platforms and associated trade-offs.

# What is the CRASH Project?

**Josh Kenway**

Fmr. Bug Bounties Research Fellow  
Algorithmic Justice League  
(Lead Co-Author)

**Camille François**

CRASH Project Co-Lead  
Algorithmic Justice League  
(Lead Co-Author)

**Dr. Sasha Costanza-Chock**

CRASH Project Co-Lead / Director of Research & Design  
Algorithmic Justice League  
(Co-Author)

**Inioluwa Deborah Raji**

AI Harms Research Fellow  
Algorithmic Justice League  
(Co-Author)

**Dr. Joy Buolamwini**

Founder & Executive Director  
Algorithmic Justice League  
(Co-Author)

# Research Collaborators & Interviewees



**Marcia Hofmann**  
Digital Rights Attorney



**Alex Rice**  
CTO, HackerOne



**Amit Elazari Bar On**  
Director, Global Policy  
Intel Corporation



**Jack Cable**  
Security Researcher



**Ryan Ellis**  
Associate Professor,  
Northeastern University



**Yuan Stevens**  
Tech Policy Research Lead,  
Ryerson University



**Rayna Stamboliyska**  
Fmr. VP – Governance,  
Yeswehack



**Lisa Wiswell-Coe**  
Fmr. Project Manager,  
Hack The Pentagon



**Marten Mickos**  
CEO, HackerOne



**Dino Dai Zovi**  
Security Researcher



**Matt Goerzen**  
Researcher,  
Data + Society

# Break It, and We'll Pay You: From Lock Picking to Breaking AI

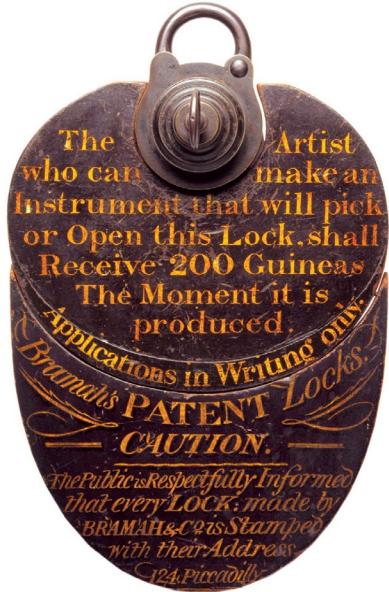


Image Credit:  
Cabinet Magazine / Science  
and Society Picture Library

**Get a bug if you find a bug.**

Show us a bug in our VRTX real-time operating system and we'll return the favor. With a bug of your own to show off in your driveway.

There's a catch, though. Since VRTX is a microprocessor operating system completely sealed in silicon, finding a bug won't be easy.

To assist along with task management and communication, memory management, and character I/O, VRTX contains over 100,000 man-hour design effort.

And since it's delivered in 4K bytes of ROM, VRTX will perform for

you the way it's performing in hundreds of real-time applications from avionics to video games. Bug free.

So, to save up to 12 months of development time and maybe give a loveable little car from the junkyard, contact us. Call (415) 326-2950, or write Hunter & Ready, Inc., 444 University Avenue, Palo Alto, California 94306.

Describe your application and the microprocessor you're using. Choose 28000, 280, 48000, or 8086 family. We'll send you a VRTX evaluation package, including timings for system calls and interrupts. And when you order a VRTX system for your application, we'll include instructions for reporting errors.\*

But don't feel bad if in a year from now there isn't a bug in your driveway.

There isn't one in your operating system either.

HUNTER & READY  
VRTX  
Operating Systems in Silicon.

\*Call or write for details. But, considering our taste in cars, you might want to accept our offer of \$1,000 cash instead. © 1983 Hunter & Ready, Inc.



Google  
hackerone



# What is algorithmic (or “AI”) harm?

**When an actor, such as an individual or an institution, uses an algorithmic system to automate classification, prediction, recommendations, scoring, or other decisions 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.**

(AJL Working Definition)



***Example:** In mid-2021, the ACLU announced they were suing the Detroit Police Department on behalf of Robert Williams, who was wrongfully arrested in part based on the failure of facial recognition technology.*

Image Credit:  
The New York Times

# Program Design Levers

<b>Target Org.</b>	Voluntary		Adversarial	
<b>Compensation</b>	Non-Monetary	Bounties	Contract	Employment
<b>Disclosure</b>	(Delayed) Full Disclosure		Coordinated Disclosure	Non-Disclosure
<b>Participation</b>	Public		Private (Invite-Only / Selective)	
<b>Program Mgmt.</b>	Platform-Managed	Mixed Management		Self-Managed
<b>Duration</b>	Ongoing		Time-Limited	
<b>Scope &amp; Access</b>	Constrained		↔	Expansive
	'Closed Box'		↔	'Open Box'

# RSA® Conference 2022

## Our Research, In Brief

**5 key research findings on bug bounties  
for algorithmic harms**



# Key Finding #1: Prepare to Include Socio-Technical Issues

- The gradual **emergence of socio-technical bounties hasn't happened in a structured way**, and **no clear best practices** have emerged yet, but the **trend is likely to continue**.
- Organizations need to ensure that they have the "**digestive systems**" (Moussouris, 2021b) in place to address identified concerns.
- **Practitioners should carefully consider the type(s) of issues that could usefully be unearthed for their organizations** (examples include data abuse, algorithmic harm, and platform exploitation).

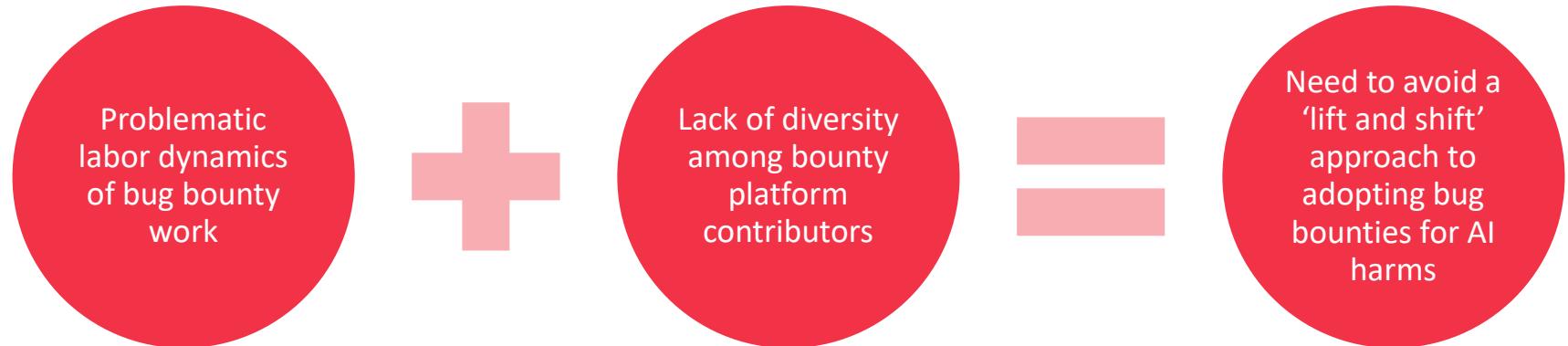
## Key Finding #2: Look Across the Lifecycle

- Bug bounties are **most impactful and efficient as one of many mechanisms** for preventing and addressing in-scope issues.
- Researchers focused on algorithmic bias and harms have emphasized a **related need to consider harms beyond a narrow focus on input data or model specification.**
- Practitioners may look to '**secure development lifecycle**' methodologies for how to structure such an approach, including **how and when bounty-like mechanisms could be most useful.**

## Key Finding #3: Nurture a Cross-Disciplinary Community of Practice

- Programs and platforms can help foster a community of practice through **education, information sharing, and trust-building**. There is a need for similar, well-curated, accessible resources to help nurture the algorithmic harms research community.
- Bounty program and platform affordances and legal terms should actively **encourage collaboration**.
- **We caution against community-building that excludes community advocates or researchers from fields outside of computer science.**

# Key Finding #4: Intentionally Develop a Diverse Community



- Platforms and programs should aim to **cultivate diverse and inclusive communities (meaningful, time-bound targets are key)**.
- Compensation mechanisms should be **fair and predictable**.
- The creation of industry standards, frameworks (e.g., impact scoring), templates, and tooling must be **inclusive and participatory**.

# Key Finding #5:

## Protect Participatory, Adversarial Research & Public Disclosure

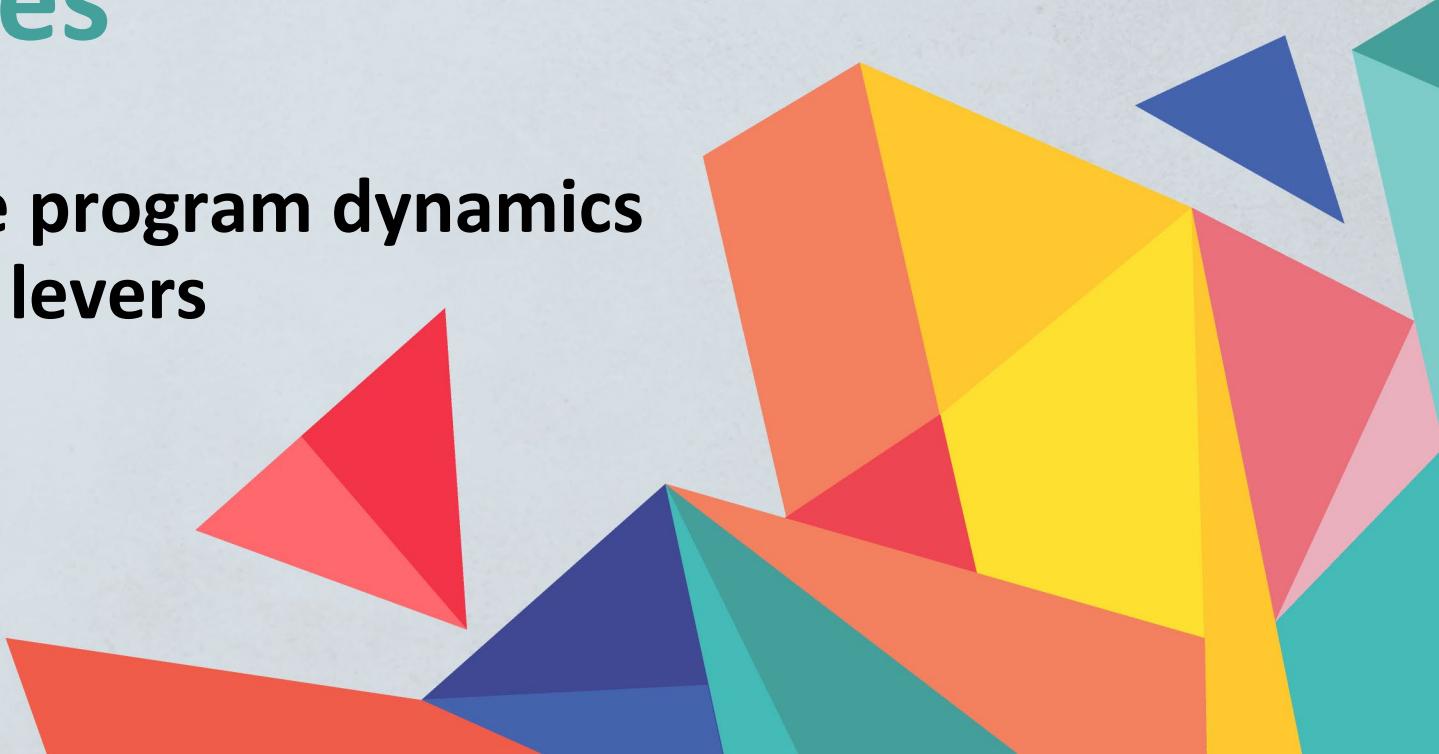
- The vulnerability disclosure space lacks successful models for scrutinizing systems that are all of the following: **participatory**, **legally safe**, **properly compensated**, and **independent**.
- **Intermediaries are incentivized to act in line with the interests of target organizations**, rather than to improve security overall.
- Discovery and disclosure of AI harms will require us to rethink:
  - **Funding of platforms** or other intermediaries
  - **Legal protection / other legal support** for researchers
  - Selection of **target organizations** and **in-scope systems**



## Flipping the script: What can be learned from our research to improve security bug bounties?

# Beyond Bug Bounties

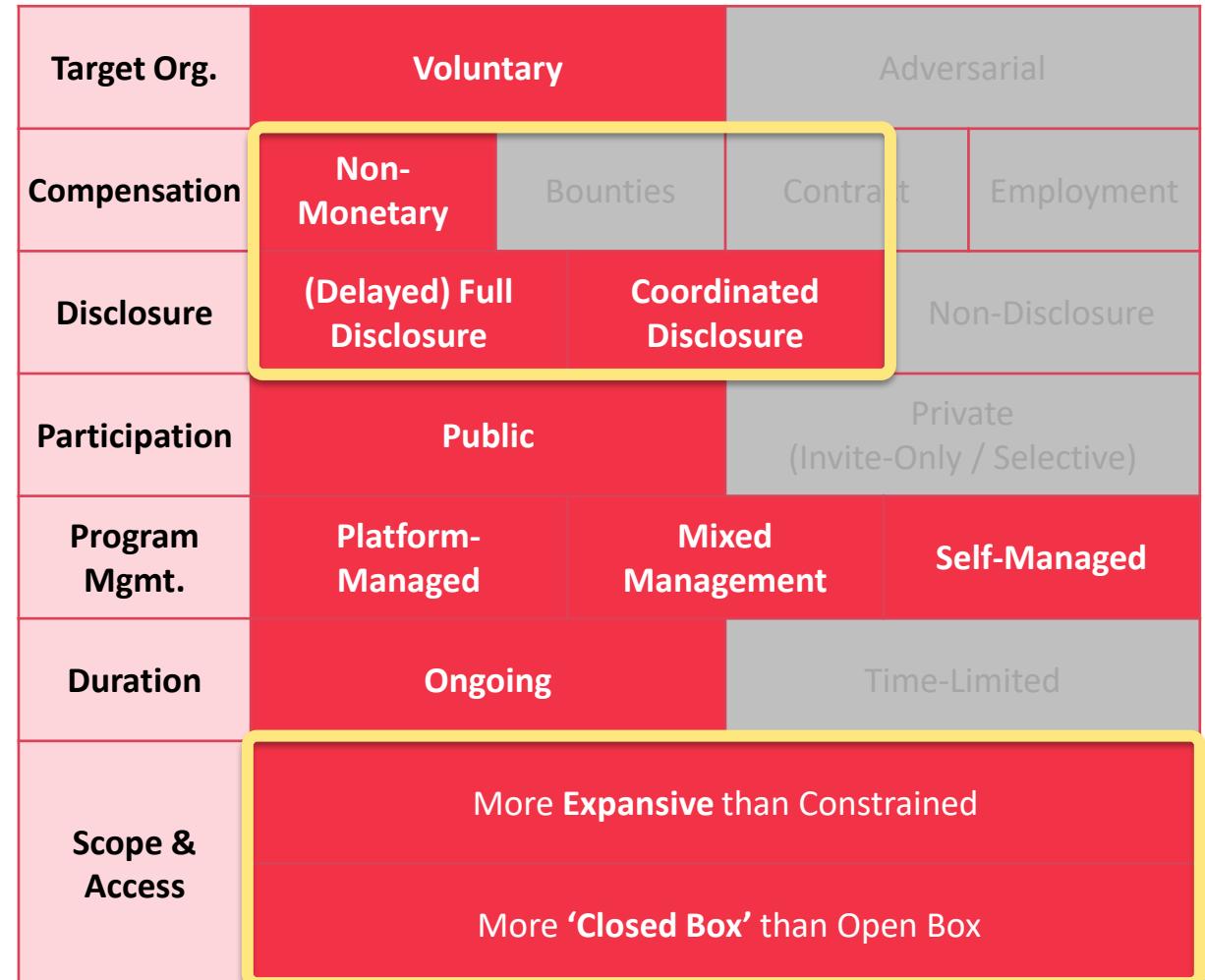
**Security bounty and disclosure program dynamics  
through the lens of our design levers**



# Vulnerability Disclosure Programs

The image shows two screenshots side-by-side. On the left is the SEGA Vulnerability Disclosure Program page, featuring a blue header with the SEGA logo and a 'Submit report' button. It displays statistics: 'Reports resolved 15' and 'Assets in scope 7'. Below this is a section titled 'Response Efficiency' stating '18 hrs Average time to first response'. A note from SEGA Europe Limited expresses appreciation for the security community's role. On the right is the CISA COORDINATED VULNERABILITY DISCLOSURE (CVD) PROCESS page, which includes a sidebar with links like 'Cybersecurity Training & Exercises', 'Cybersecurity Summit 2020', and 'Protecting Critical Infrastructure'. The main content area describes the CVD program's goal of coordinating remediation and public disclosure of vulnerabilities.

## "Typical" Vuln Disclosure Program Design



# Self-Managed Bug Bounties

VentureBeat

**After paying over \$4M in bounties since 2010, Google expands bug bounty program to its Android and iOS apps**

Emil Protalinski  
@EPro

January 30, 2015

PRESS RELEASES

**Mozilla Foundation announces first payments of Security Bug Bounty Program, further strengthens browser security**

September 14, 2004

MOUNTAIN VIEW, Calif. – September 14, 2004 – One month after announcing its Security Bug Bounty Program, the Mozilla Foundation is showing the first positive results from this initiative to enlist the help from the open source developer community to make its browsers even more secure. The Mozilla Foundation today released updates to its Firefox and Mozilla 1.7 browsers and Thunderbird email client that include a number of security enhancements and address several potential security vulnerabilities, taking a proactive leadership role in protecting Internet users from malicious attacks.

"Typical" Self-Managed Bounty Program Design

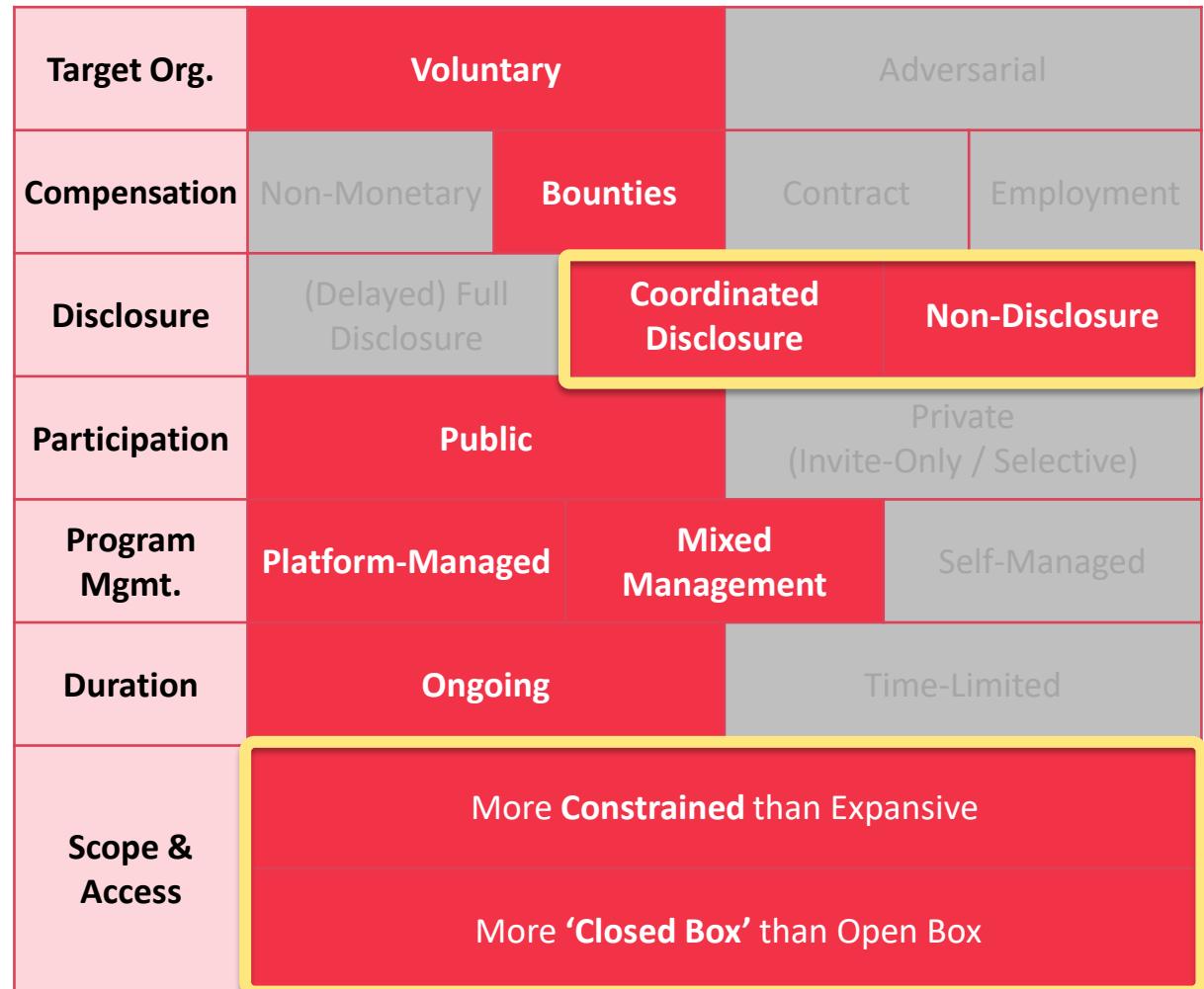
Target Org.	Voluntary		Adversarial	
Compensation	Non-Monetary	Bounties	Contract	Employment
Disclosure	(Delayed) Full Disclosure		Coordinated Disclosure	Non-Disclosure
Participation	Public		Private (Invite-Only / Selective)	
Program Mgmt.	Platform-Managed	Mixed Management	Self-Managed	
Duration	Ongoing		Time-Limited	
Scope & Access	More Constrained than Expansive More 'Closed Box' than Open Box			

# Platform Bug Bounties – Public

**Nintendo**

- Submit report
- Bug Bounty Program Launched on Dec 2016
- Includes retesting
- Bounty splitting enabled
- Reports resolved: 84
- Assets in scope: -
- Average bounty: \$1k-\$2k
- Policy, Hacktivity, Thanks, Updates (1), Collaborators
- Policy: Nintendo's goal is to provide a secure environment for our customers so that they can enjoy our games and services. In order to achieve this goal, Nintendo is interested in receiving vulnerability information that researchers may discover regarding Nintendo's platforms. Currently, in the context of the HackerOne program, Nintendo is only interested in vulnerability information regarding the Nintendo Switch™ family of systems and is not seeking vulnerability information regarding other Nintendo platforms, network service, or server-related information.
- Below are examples of:
  - Piracy, including:
    - Game application
    - Copied game application
  - Cheating, including:
    - Game application
    - Save data modification
  - Dissemination of information
- USAA
- Submit report
- Program details: Announcements, Hall of Fame
- USAA appreciates and supports engagement with security community when potential security vulnerabilities in our digital assets are reported to us in accordance with Responsible Disclosure policy.
- Ratings/Rewards: For the initial prioritization/rating of findings, this program will use the Bugcrowd Vulnerability Rating Taxonomy. However, it is important to note that in some cases a vulnerability priority will be modified due to its likelihood or impact. In any instance where an issue is downgraded, a full, detailed explanation will be provided to the researcher - along with the opportunity to appeal, and make a case for a higher priority.
- Scope and rewards: High Priority Targets, In scope
- Latest hall of famers: [Circular icons]

## "Typical" Public Platform Bounty Program Design



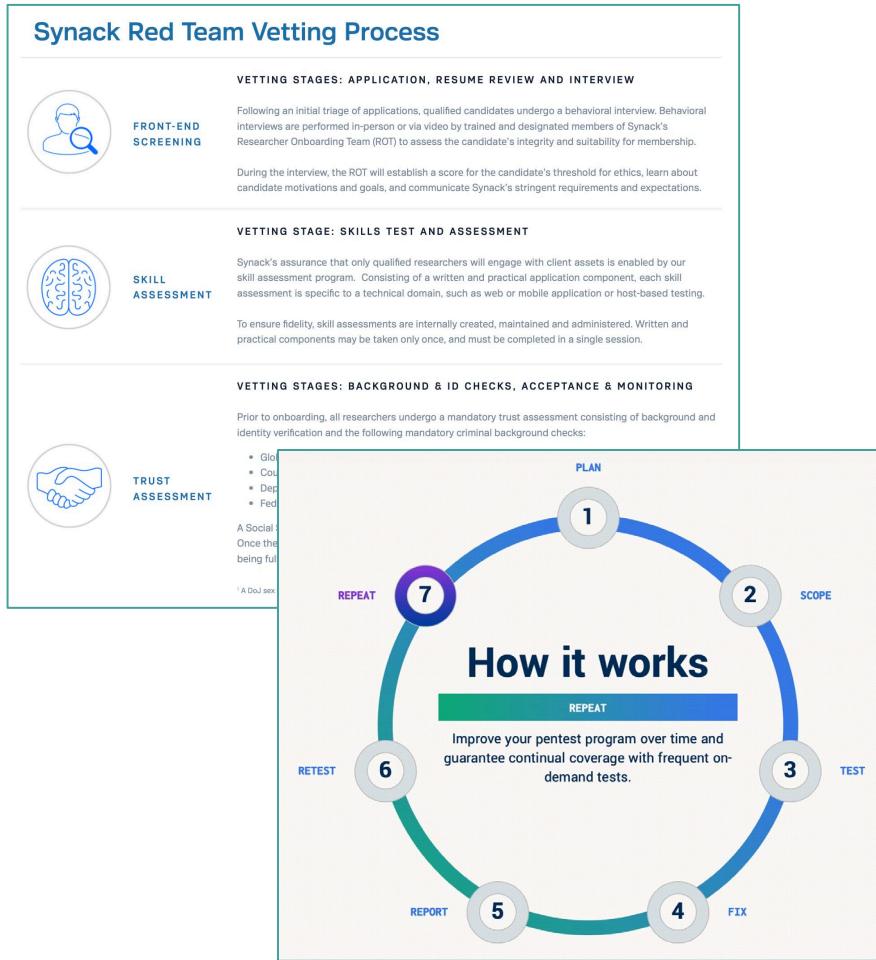
# Platform Bug Bounties – Private



**“Typical” Private Platform Bounty Program Design**

<b>Target Org.</b>	Voluntary		Adversarial	
<b>Compensation</b>	Non-Monetary	Bounties	Contract	Employment
<b>Disclosure</b>	(Delayed) Full Disclosure	Coordinated Disclosure	Non-Disclosure	
<b>Participation</b>	Public		Private (Invite-Only / Selective)	
<b>Program Mgmt.</b>	Platform-Managed	Mixed Management	Self-Managed	
<b>Duration</b>	Ongoing		Time-Limited	
<b>Scope &amp; Access</b>	More Constrained than Expansive More ‘Open Box’ than Closed Box			

# 'Crowdsourced Pentesting'



## 'Crowdsourced Pentesting' Program Design

Target Org.	Voluntary	Adversarial	
Compensation	Non-Monetary	Bounties Contract Employment	
Disclosure	(Delayed) Full Disclosure	Coordinated Disclosure	Non-Disclosure
Participation	Public	Private (Invite-Only / Selective)	
Program Mgmt.	Platform-Managed	Mixed Management	Self-Managed
Duration	Ongoing	Time-Limited	
Scope & Access	More Constrained than Expansive		
	More 'Open Box' than Closed Box		

# Industry Bounties



## Industry Bounties Program Design

<b>Target Org.</b>	Voluntary			<b>Adversarial</b>		
<b>Compensation</b>	Non-Monetary	<b>Bounties</b>	Contract	Employment		
<b>Disclosure</b>	<b>(Delayed) Full Disclosure</b>		Coordinated Disclosure	Non-Disclosure		
<b>Participation</b>	<b>Public</b>		<b>Private (Invite-Only / Selective)</b>			
<b>Program Mgmt.</b>	<b>Platform-Managed</b>	Mixed Management	Self-Managed			
<b>Duration</b>	<b>Ongoing</b>		Time-Limited			
<b>Scope &amp; Access</b>	More <b>Expansive</b> than Constrained					
	More ' <b>Closed Box</b> ' than Open Box					

# Open-Source / Non-Profit Bug Bounties

**We fund open source security.**  
We pay security researchers for finding vulnerabilities in any GitHub repository and maintainers for fixing them.

Join the fight | Come talk

90% of users Got their first CVE 1.7 CVEs Avg. per user 3.6 bounties per user Avg. monthly winnings

**Internet Bug Bounty**  
The Internet Bug Bounty rewards security research into vulnerabilities impacting Open Source Software Projects.

<https://www.hackerone.com/internet-bug-bounty>

Reports resolved 693 Assets in scope 17 Average bounty \$500-\$1k

## Open-Source / Non-Profit Bounty Design

Target Org.	Voluntary		Adversarial						
Compensation	Non-Monetary	Bounties	Contract	Employment					
Disclosure	(Delayed) Full Disclosure	Coordinated Disclosure	Non-Disclosure						
Participation	Public		Private (Invite-Only / Selective)						
Program Mgmt.	Platform-Managed	Mixed Management	Self-Managed						
Duration	Ongoing	Time-Limited							
Scope & Access	More <b>Expansive</b> than Constrained								
	More ' <b>Open Box</b> ' than Closed Box								

# What's Old Could Be New Again

Our findings suggest that independent bounties could usefully complement the status quo



# Independent Adversarial Bounties (IABs)

- The concept of an adversarial bug bounty with non-profit intermediaries was floated in the late 1990s / early 2000s
- However, this model was mostly superseded by vendor and operator-driven bounties and VDPs
- For-profit intermediaries don't resolve longstanding incentive misalignment around public disclosure, prioritization, and scope
- Our research illustrates how such a design niche could be an effective complement to existing bug bounties and VDPs

# IAB Design Overview

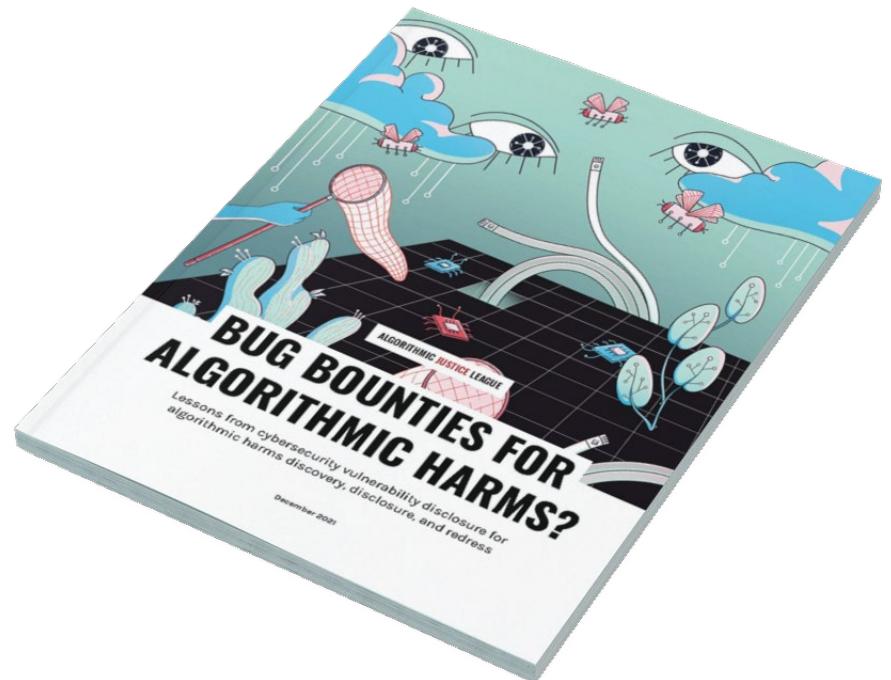
Target Org.	Voluntary		Adversarial	
Compensation	Non-Monetary	Bounties	Contract	Employment
Disclosure	(Delayed) Full Disclosure	Coordinated Disclosure		Non-Disclosure
Participation	Public		Private (Invite-Only / Selective)	
Program Mgmt.	Platform-Managed	Mixed Management		Self-Managed
Duration	Ongoing		Time-Limited	
Scope & Access	Constrained or Expansive, as needed Closed Box			

# IAB Actionable Recommendations

What is needed for this model to be realized and successful?

Governments	Non-profit intermediaries	Vendors / operators	Bounty platforms
<ul style="list-style-type: none"><li>• Public funding of coordinating non-profits</li><li>• Legal guidance + reforms where needed to protect participating researchers and facilitate access to closed systems</li></ul>	<ul style="list-style-type: none"><li>• Independent, priority-based selection of target organizations and systems</li><li>• Flexible models for researcher participation (bounties / grant-based)</li></ul>	<ul style="list-style-type: none"><li>• Broadening scope of existing bounty programs / VDPs to minimize foreseeable impact</li><li>• Adaptation and development of internal processes and policies</li></ul>	<ul style="list-style-type: none"><li>• Licensing or donation of coordination infrastructure</li><li>• Vetting of researcher experience / trustworthiness</li></ul>

Read the report at:  
[ajl.org/bugs](http://ajl.org/bugs)



Thank you!