RSAConference2022
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.

RSAConference2022

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

Google

hackerone

iDEFENSE℠
The Power of Intelligence

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

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 ↔ Expansive			
	‘Closed Box’ ↔ ‘Open Box’			

RSAConference2022

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).

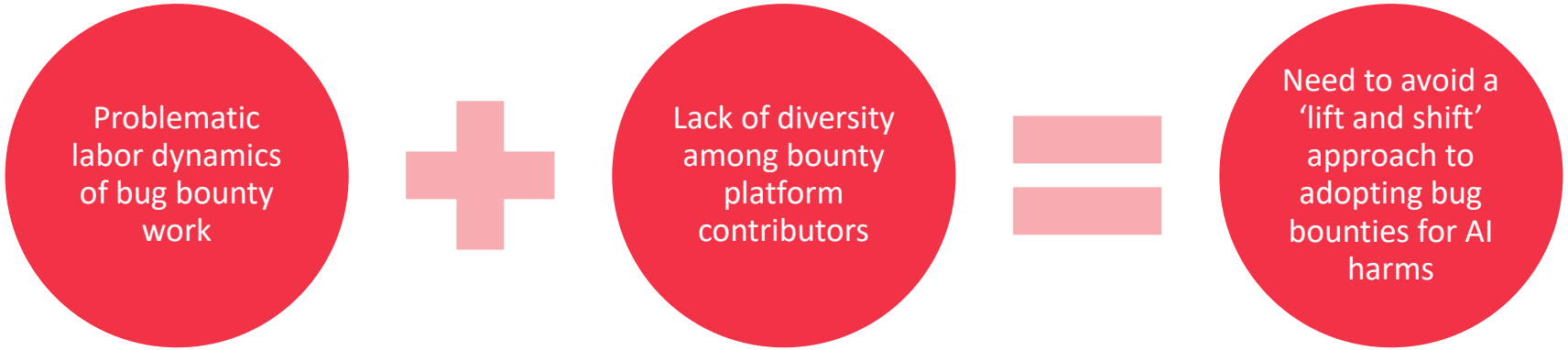
Key Finding #2: Look Across the Lifecycle

- 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-bounded 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**

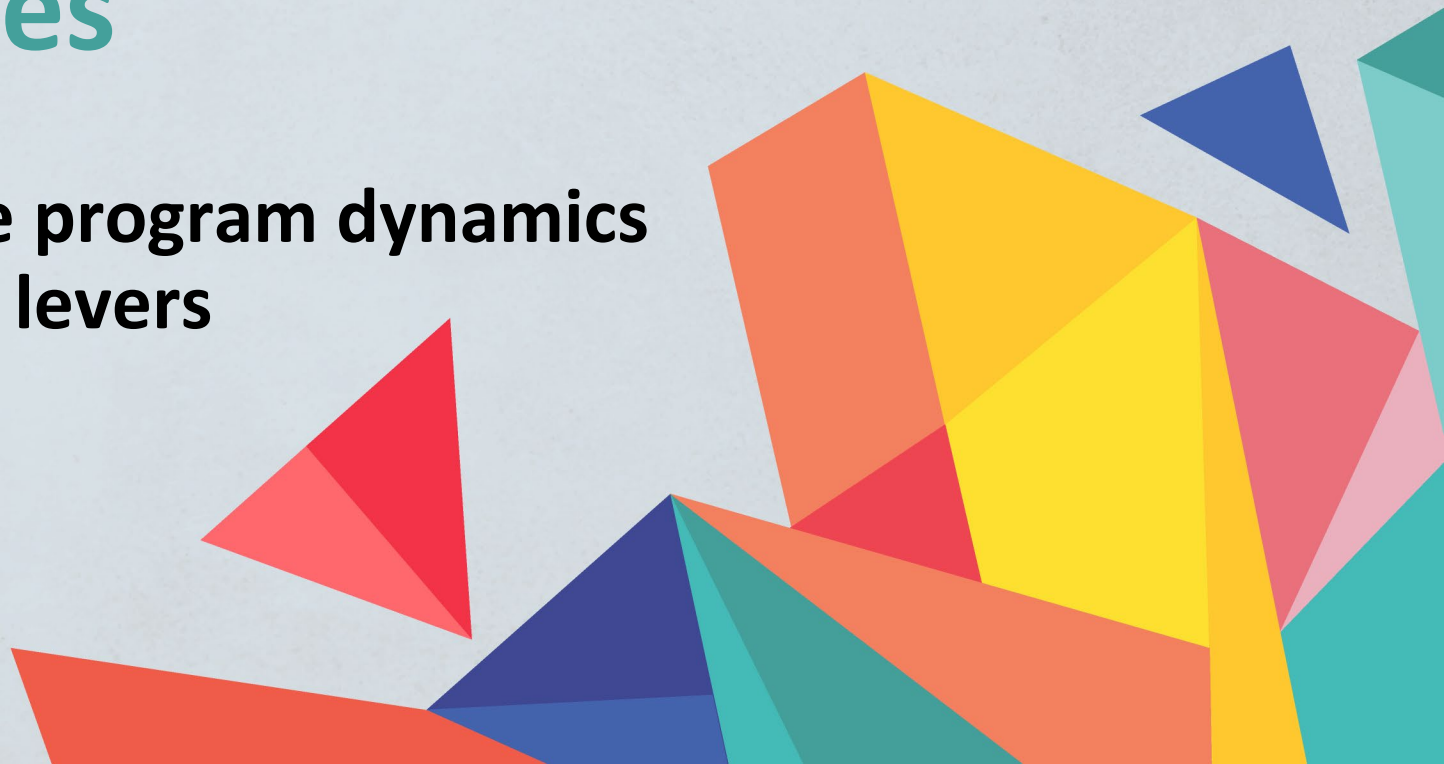
Flipping the script:

What can be learned from our research to improve security bug bounties?

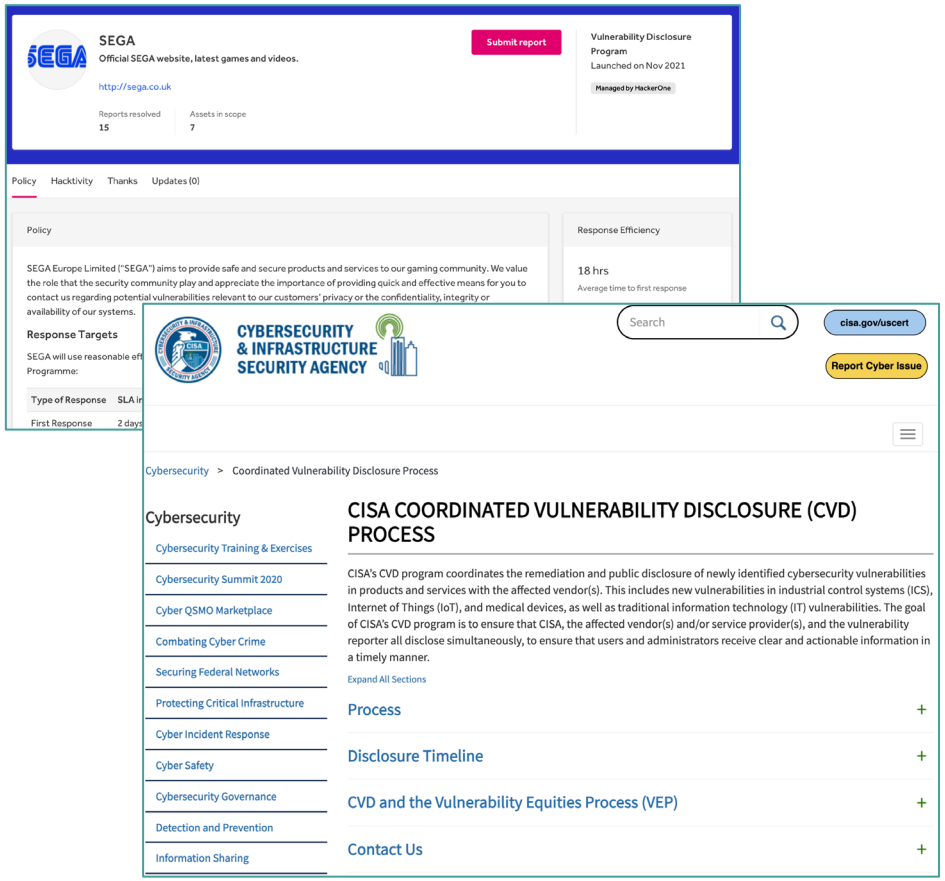
RSAConference2022

Beyond Bug Bounties

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



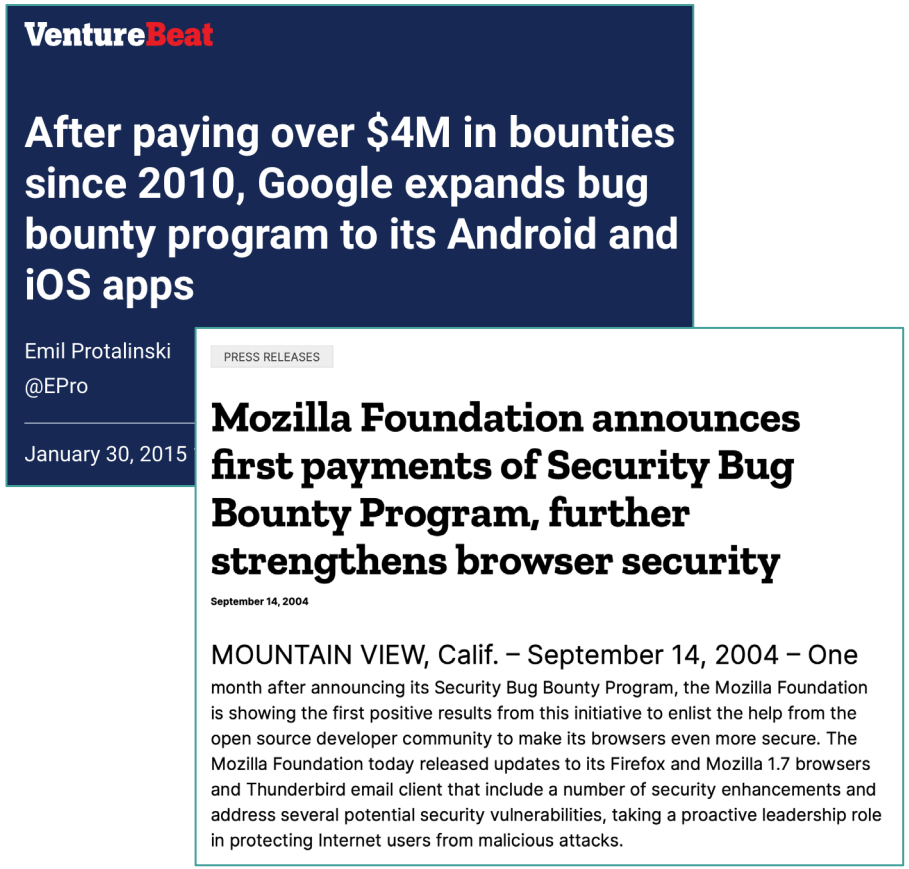
Vulnerability Disclosure Programs



"Typical" Vuln Disclosure 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 **Expansive** than Constrained More **'Closed Box'** than Open Box	

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

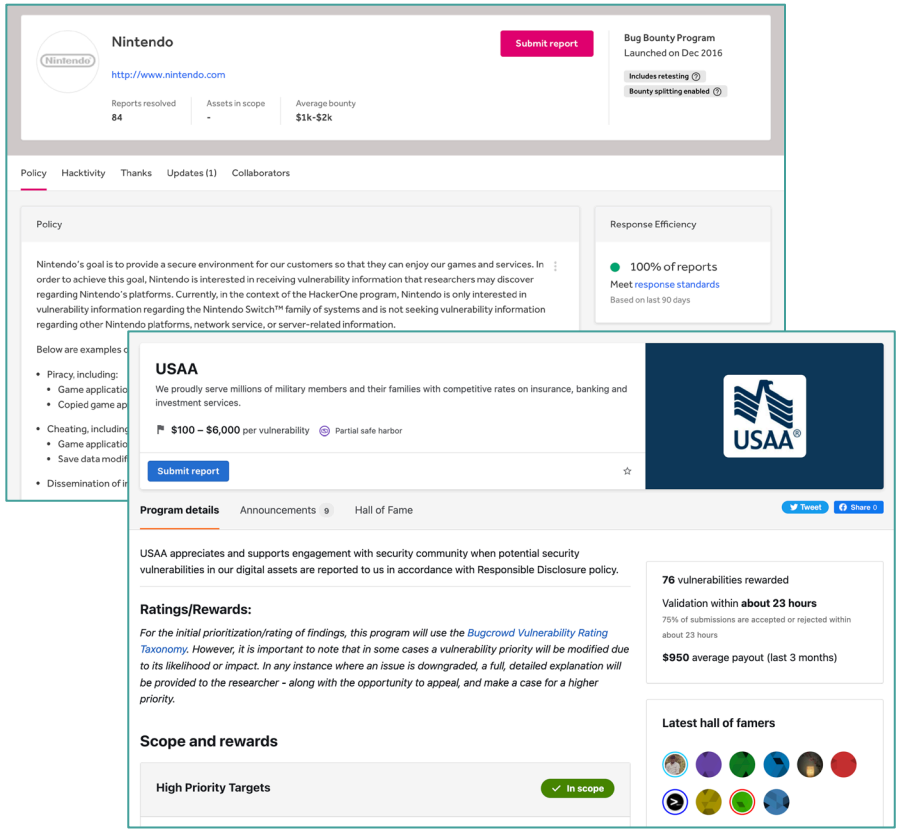
September 14, 2004

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



"Typical" Public Platform 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 – Private

"Typical" Private Platform 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 **'Open Box'** than Closed Box			

‘Crowdsourced Pentesting’

Synack Red Team Vetting Process

FRONT-END SCREENING

VETTING STAGES: APPLICATION, RESUME REVIEW AND INTERVIEW

Following an initial triage of applications, qualified candidates undergo a behavioral interview. Behavioral interviews are performed in-person or via video by trained and designated members of Synack's Researcher Onboarding Team (ROT) to assess the candidate's integrity and suitability for membership.

During the interview, the ROT will establish a score for the candidate's threshold for ethics, learn about candidate motivations and goals, and communicate Synack's stringent requirements and expectations.

SKILL ASSESSMENT

VETTING STAGE: SKILLS TEST AND ASSESSMENT

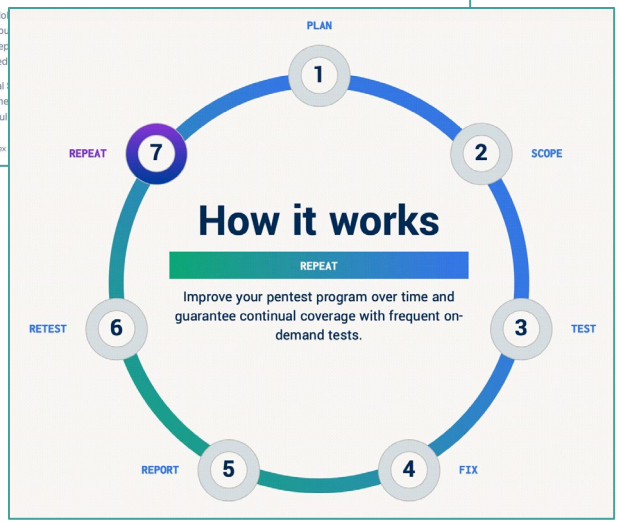
Synack's assurance that only qualified researchers will engage with client assets is enabled by our skill assessment program. Consisting of a written and practical application component, each skill assessment is specific to a technical domain, such as web or mobile application or host-based testing.

To ensure fidelity, skill assessments are internally created, maintained and administered. Written and practical components may be taken only once, and must be completed in a single session.

TRUST ASSESSMENT

VETTING STAGES: BACKGROUND & ID CHECKS, ACCEPTANCE & MONITORING

Prior to onboarding, all researchers undergo a mandatory trust assessment consisting of background and identity verification and the following mandatory criminal background checks:



‘Crowdsourced Pentesting’ Program Design

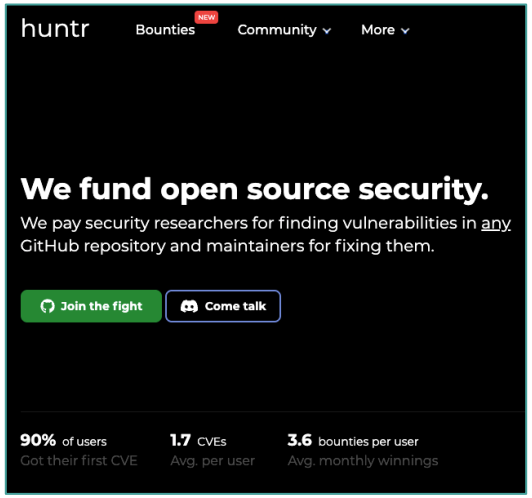
Target Org.	**Voluntary**	Adversarial	
Compensation	Non-Monetary	**Bounties**	**Contract**	Employment
Disclosure	(Delayed) Full Disclosure	Coordinated Disclosure	**Non-Disclosure**
Participation	Public	**Private (Invite-Only / Selective)**	
Program Mgmt.	**Platform-Managed**	Mixed Management	Self-Managed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

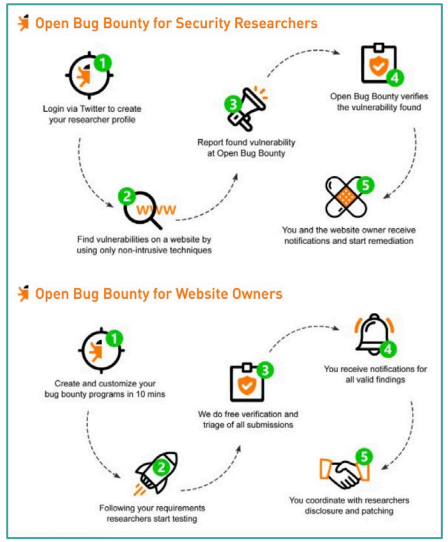
Industry Bounties

Industry Bounties Program Design

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 **Expansive** than Constrained			
	More **'Closed Box'** than Open Box			

Open-Source / Non-Profit Bug Bounties





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 **Expansive** than Constrained			
	More **'Open Box'** than Closed Box			

RSAConference2022

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

- Public funding of coordinating non-profits
- Legal guidance + reforms where needed to protect participating researchers and facilitate access to closed systems

Non-profit intermediaries

- Independent, priority-based selection of target organizations and systems
- Flexible models for researcher participation (bounties / grant-based)

Vendors / operators

- Broadening scope of existing bounty programs / VDPs to minimize foreseeable impact
- Adaptation and development of internal processes and policies

Bounty platforms

- Licensing or donation of coordination infrastructure
- Vetting of researcher experience / trustworthiness

Read the report at: ajl.org/bugs

Thank you!