On the Acceptance by Code Reviewers of Candidate Security Patches Suggested by Automated Program Repair Tools

Empirical Software Engineering Journal (RR ESEM'22)

APR tools in a Nutshell and Challenges

- APR tools alleviate the manual effort involved in fixing bugs by suggesting patches to automatically fix them.
- Patches identified by APR tools may have passed all automatic tests and still be semantically incorrect (e.g. Liu et al. JSS 2021)
- Change-based code review problem (e.g. Braz et al. ICSE 2022)

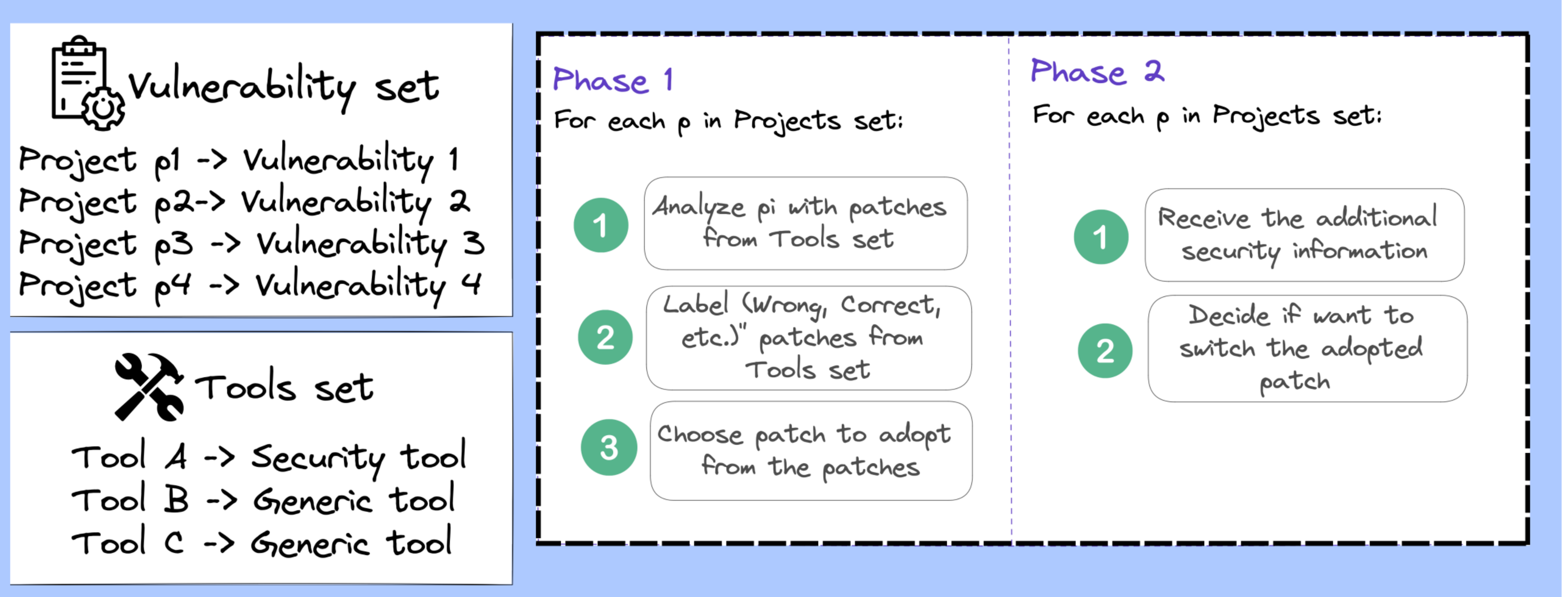
VU Vrije Universiteit Amsterdam — Università di Trento

Theseus — Dutch Research Council (NWO) Theseus (NWA.1215.18.006)

AssureMOSS — H2020 Program AssureMOSS (Grant n.952647)

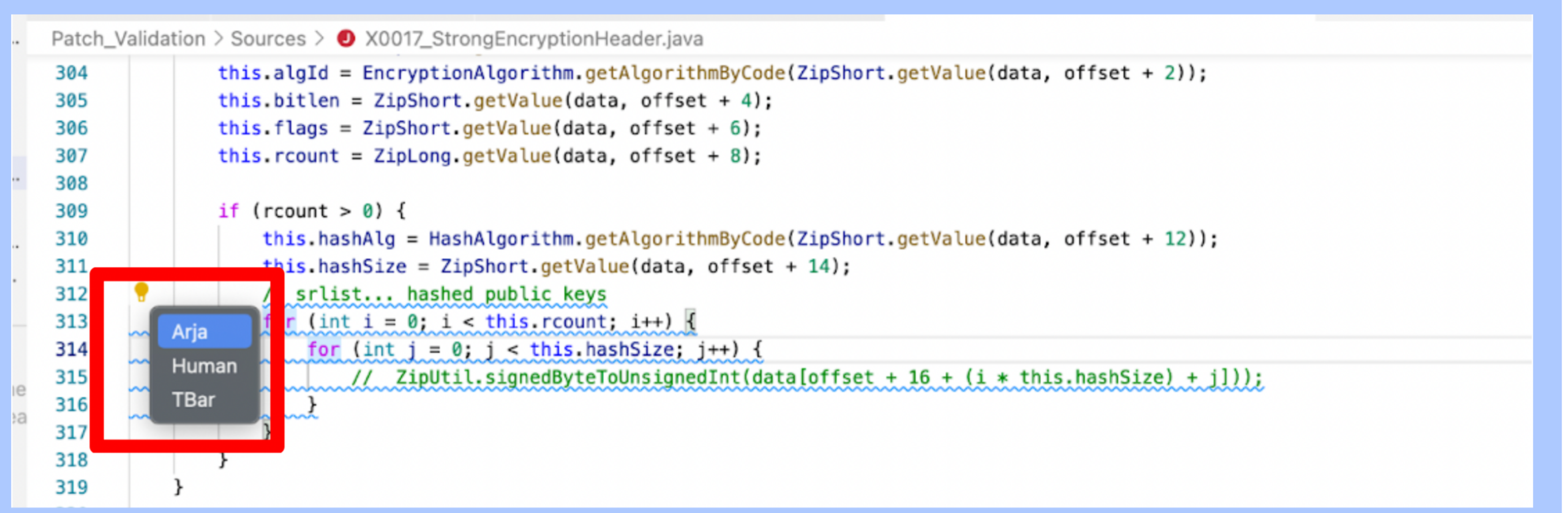
SEC4AI4SEC — HORIZON EUROPE Sec4AI4Sec (Grant n.952647)

Execution Plan



Automated Vulnerability Repair Task

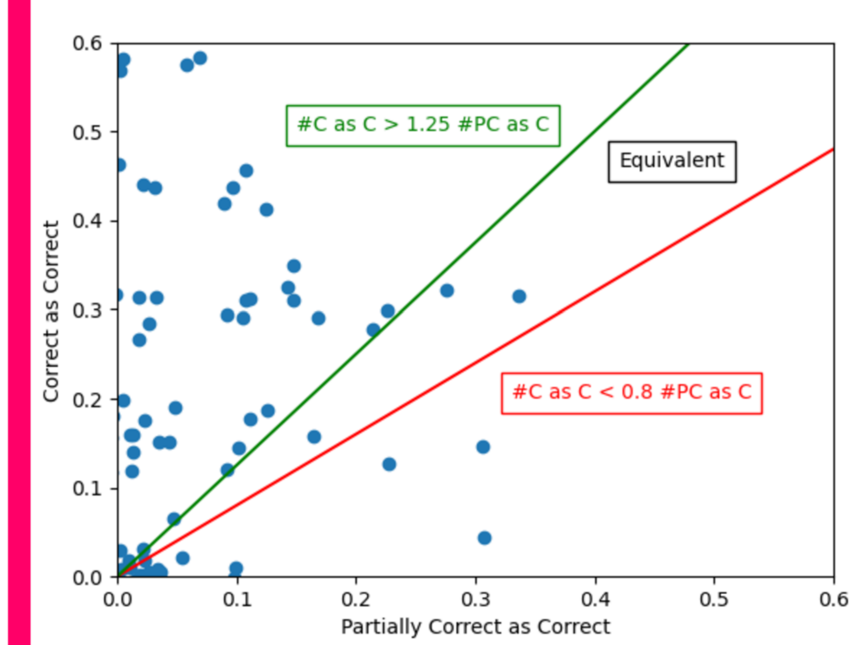
Is this patch (a) Correct (b) Partially Correct or (c) Wrong?



RQ1

Will human code reviewers be able to discriminate between correct and wrong security patches submitted by the APR tools?

- **It is EASIER to identify WRONG patches than CORRECT patches.**
- **Correct patches are not confused with partially correct patches**
- **Patches from APR4Sec are adopted more often than patches suggested by generic APR tools.**



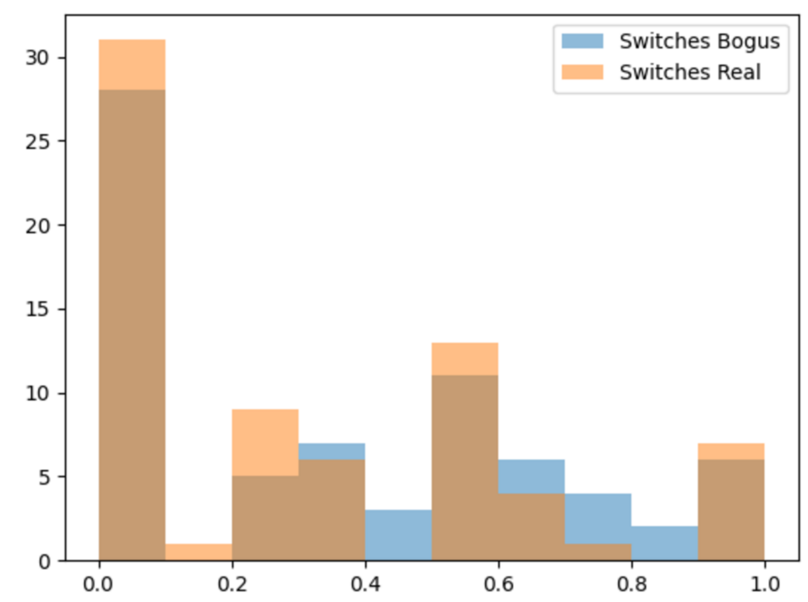
The green and the red lines correspond to the values X*0.8 < Y < X*1.25 where X is partially correct patches identified as correct patches, and Y is correct patches identified as correct patches.

Correct patches (Y) are even higher than the 125% value of the partially correct patches (X). The coordinates of data points have been slightly randomized by an offset in the range [-0.01, 0.01].

RQ2

Will code reviewers' decisions be actually influenced by knowing that some patches come from a specialized security tool?

- **Not enough evidence to conclude that `bogus' security claims are either indistinguishable or different from `true security' claims.**
- **Knowing a patch is from a security tool INCREASES the chances of adoption irrespective of correctness.**



On the X axes there is the proportion of actual switches with respect to potential switches available to the participants, and on Y axes there is the frequency.

As one can see there is a higher proportion for the bogus treatment rather than the real treatment.

We can notice a large number of zeros, which it represents no switches.

Summary

- Are humans able to recognize the semantic correctness (passed all automatic tests) of APR tools patches?
  - Correct vs Partially Correct vs Wrong
  - Is it biased knowing the APR tool is designed for security?
- Perform a controlled experiment with humans
  - 72 master's students
  - 7 CVEs and 7 APR tools (Generic and Security)
- Possible collaborations: (1) experiment replication (2) and more APR tools to test

AURORA PAPOTTI[2] a.papotti@vu.nl

RANINDYA PARAMITHA[1] ranindya.paramitha@unitn.it

FABIO MASSACCI[1,2] fabio.massacci@ieee.org

(1) Università di Trento, Italy; (2) Vrije Universiteit Amsterdam, The Netherlands