Teaching Adversarial Thinking by Having Students Circumvent Exam Rules

Matthew Bowden, Tom Chothia, Anna Clee, Sam Collins(✉),
Jacqueline Henes, and David Oswald

University of Birmingham, Birmingham, UK
{mxb1143,axc1017,sxc1327,jkh703}@student.bham.ac.uk,
{t.chothia,d.f.oswald}@bham.ac.uk

Abstract. This paper describes a live exercise in adversarial thinking, deployed as part of an advanced undergraduate module: *Security of Real World Systems*. Delivered as a mock exam, students must answer a single question given two weeks in advance: write the first sixty digits of π. We present a student attacker model, from which we construct a set of exam rules, all containing deliberate oversights. Students must analyse the rules for vulnerabilities, develop a solution, and successfully deploy it in a mock exam invigilated by the module team. We describe how we handled communication with students, invigilation, and marking such an exercise. We run an educational evaluation, categorise student solutions, and gauge student opinions on the exercises from feedback on the module, as well as a direct focus group. Finally, we discuss some more salient lessons from the exercise and how we might address them as we take our work forward.

1 Introduction

Teaching the practical skills required in the field of cybersecurity is innately difficult - the field does not lend itself well to pure bookwork, and exercises can often be lengthy in their setup and difficult in their deployment. Additionally, a balance must be found between theory, practical skills, and fostering the creative approaches that the field demands in students.

In this work, we describe an exercise in adversarial thinking that we ran as part of the 3rd year B.Sc. course *Security of Real-World Systems* at the University of Birmingham, UK. This is an advanced cybersecurity module that covers topics such as buffer overflows, ROP attacks, side-channel analysis, and protocol attacks. A recurring theme - and key learning objective - of the course is that understanding the *rules* of a system well often makes it possible to discover exploits that were not foreseen when it was created. For example, understanding the layout of the stack and calling convention used in x86_64 allows for return-to-libc and ROP attacks, and understanding the instruction-dependent timing and power usage of CPUs reveals a side-channel attack on AES.

Based on an idea by Conti and Caroland [3], we set a mid-term "exam" that asked students to write out the first sixty digits of π. This exam-like assignment

P. Legg et al. (Eds.): CSE 2024, LNNS 1213, pp. 75–95, 2024.
https://doi.org/10.1007/978-3-031-77524-6_5

had a carefully crafted set of rules that would be enforced as written. While initially seeming impossible to solve, the exam rules contained enough loopholes for the student to complete the task. Conti and Caroland encourage their students to cheat and break the rules [3]; conversely the objective of our exercise is for the students to analyse the rules; find weaknesses in them; and exploit them; this is in line with the key learning objectives of the course.

We ran the mid-term exam by replicating the format of our institution's exam procedure, with the exercise worth a token 2% of the overall course mark. The students demonstrated a wide range of creative solutions with a mix of approaches both expected and unexpected. Student feedback on the test indicated that this was very popular, and that the learning objective was well understood.

In summary, the contributions of this work are:

- A student attacker model designed to promote outside-the-box analysis and adversarial thinking.
- A framework for developing mock exam rules with exploitable oversight, as well as our ruleset built to match the student attacker model.
- Guidance on communication with students before such an exercise, invigilation, and assessment .
- An educational evaluation of our results.
- Details of practical lessons learned, with guidance for other institutions who may wish to deploy a similar exercise.

Our ruleset can be found in Appendix and further details can be found at https://www.cs.bham.ac.uk/~tpc/ExamExercise/.

2 Background

2.1 Live Security Exercises

Cybersecurity lies apart from other disciplines in the fields of science and engineering due to its heavily adversarial nature. Progress in cybersecurity is not just driven by core count and computational complexity, but by new methods to obfuscate, outmaneuver, and trick adversaries. Therefore, live security exercises and gamified challenges have become popular in the field [4], chief among them being Capture The Flag (CTF) exercises. Past research [2] has neatly classified CTFs into the following two categories:

- **Adversarial (Attack/Defence) -** whereby participants/teams must protect their vulnerable services while exploiting the other teams vulnerable services.
- **Jeopardy Style -** where there are a wider range of standalone attack-only challenges which doesn't require/support interaction between participants.[1]

[1] Davis et al. [4] suggests that if you want a defence-only CTF, you need only host something online and wait for it to be attacked.

This distinction is important, as there has been further research into the benefits and respective drawbacks of each style of CTF. Davis et al. [4] suggests that students see the defence side of adversarial CTFs to be more of an annoyance, suggesting a mirror to real life sports where the star athletes on a team tend to be the ones who score points. This has given rise to the Jeopardy-style CTF, which are now far more common, especially for learning and assessment at a more beginner to intermediate level.

Other forms of live security exercise exist and have been successfully deployed both in and out of education. Many share common traits with more classic CTF challenges; we discuss this further in the following section (sect 3).

3 Related Work

Within the specific context of education, research shows the value of gamified live security exercises. The United States Air Force Academy (USAFA) found increased motivation in their undergraduate students when completing an attack-only CTF challenge [1] deployed in stages through the course, when contrasted against their regular assignments. Students were found to spend additional time researching complex and novel solutions to more difficult challenges, matched with an increase in enthusiasm.

A similar approach has been demonstrated in an alternate educational setting at UC Santa Barbara by Vigna [7] in teaching network security. Three challenges were deployed on the course, similarly to Carlisle, Chiaramonte, and Caswell [1], but as standalone exercises rather than a continuously evolving CTF. Importantly they find that more disruptive attacks, in their case Denial of Service on a network, interfere with the learning objectives of the exercise for other students. We take steps to avoid disruptive interference in our exercise, discussed in Sect. 4.3.

Ensafi, Jacobi, and Crandall [5,6] deployed their live security exercise as a variant of the social deduction game Werewolf on a mix of Undergraduate and Postgraduate students. Here students connect to the game server via SSH and are expected to exploit vulnerabilities in the server to deduce the identities of their classmates. They found mixed results; students exceeded expectations regarding the sophistication of attacks and demonstrated a high level of enthusiasm for the exercise. However, they struggled to make the game fair, with the students assigned the werewolf role having little chance to succeed when faced with more decisive tactics.

Making students cheat in a simulated exam has also been explored by US Cyber Command [3]. Deployed on a small group of students to encourage the development of a more 'devious' and adversarial approach to problem-solving, students were allowed complete freedom in how they cheated and were encouraged to directly break the rules given. This approach differs slightly from our own, where we allow cheating, but encourage bypassing the rules rather than breaking them; this encourages students to perform an analysis of the system.

4 Running the Exam

4.1 Attacker Model

When designing the exam, we first define a student attacker model as a framework to consider when constructing the rules. First, students have full knowledge of system rules as well as the system's environment from when the exam is announced. They are then able to use the allotted preparation time to analyse the system, consider vulnerabilities, and develop an attack. The objective is also announced ahead of the exam, but is of less consequence than the rules or environment. A key property of this model is that it is as liberal and open ended as possible within the bounds of safety, the law, and disruption to other students. Students may attempt any attack, including social engineering and insider attacks[2]. We only had one attempted phishing attempt before the exam, shown in Fig. 1.

Worth considering is the factors that students do not have complete control over; for example, the environment the exercise takes place in. Telling the students of the environment (e.g. the exact room number in a building) in advance gives information that can be considered as part of their design, however relevant setup and precautions should be taken so that the exercise difficulty is not undermined. In our case, many suspicious items were removed from the examination room shortly before the start of the exam, as was said would be done in the rules. In one case, a student intended to use the computer at the lectern to retrieve the digits of π; in actuality this computer was unavailable for use at the time and they were unable to use this solution.

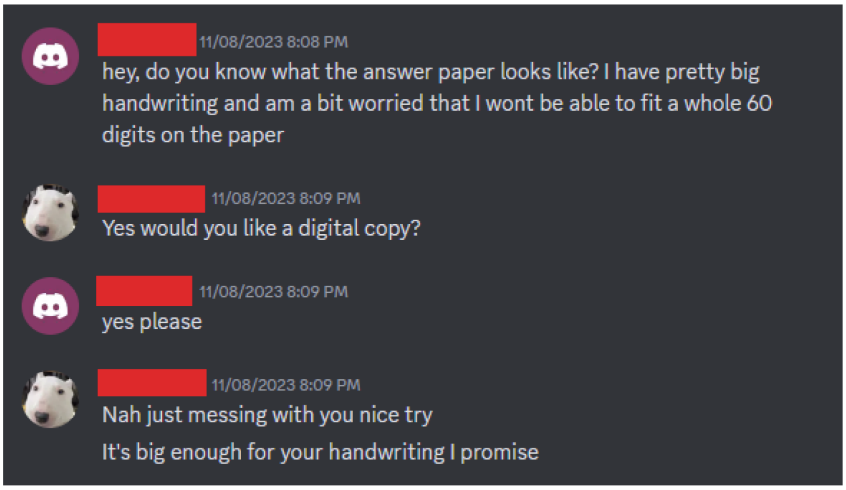
4.2 Exam Rules

The purpose of this exam was not to directly assess students understanding of specific course content, but to encourage them to engage with the target system as an adversary might. The exam rules should therefore support the abstract problem- solving abilities that cybersecurity work can demand, and in particular encourage novel solutions. To match the desired student attacker model, we released the set of exam rules and venue information two weeks before the exam, providing time for analysis and preparation. Alongside the rules we also released the objective of the exam, a single question: to list the first sixty digits of π.

Given the standard expectations of exams at the University of Birmingham[3], we believed students may need reassurance regarding the behaviour we expected of them as well as the consequences (or lack thereof) if they were

[2] We decided that attempting to convince an invigilator to render assistance is permissible, but we agreed that invigilator responses to such attempts should be pre-agreed, uniform, and fair.

[3] Please refer to https://intranet.birmingham.ac.uk/as/registry/exams/rules/rules-and-regulations.aspx for standard exam rules.



[redacted] 11/08/2023 8:08 PM
hey, do you know what the answer paper looks like? I have pretty big handwriting and am a bit worried that I wont be able to fit a whole 60 digits on the paper

[redacted] 11/08/2023 8:09 PM
Yes would you like a digital copy?

[redacted] 11/08/2023 8:09 PM
yes please

[redacted] 11/08/2023 8:09 PM
Nah just messing with you nice try
It's big enough for your handwriting I promise

Fig. 1. The only social engineering attack we experienced, attempted a few days before the exam via Discord with a TA.

caught “cheating”. Released as a part of the rule-set was a disclaimer paragraph reassuring students that circumventing the rules was encouraged.

> The rules of this “exam” will be enforced strictly and to the letter. There are no other rules. All the normal expectations of good, honourable behaviour that are required for the other exercises on this module, and other modules, are not required here. The normal description of academic misconduct and examination irregularities in the student handbook do not apply. If the rules outlined above are deemed to be broken, you will only fail this test, it will not be treated as a genuine case of an examination irregularity.

We encourage students not to cheat by breaking rules, but rather to circumvent them. We therefore disincentivise breaking the rules by adding the risk of failing the test if caught doing so. There is however a balance between there being a consequence of being caught and the severity of the consequences students are afraid of. We want them to have the confidence to attempt circumventing the rules, so we make sure they understand this will in no circumstance be treated seriously, beyond losing the marks of this exam, and there will be no disciplinary actions if caught. This is a complicated line to balance: students will need explicit reassurances as it goes against everything they have experienced so far in higher education. We discuss the importance of good communication with students and how we went about achieving this in more depth in Sect. 4.3.

We do not believe that any specific rules are necessary to run this exercise successfully; instead, rules should be designed in a way that invites students to find vulnerabilities in the system which they can the exploit. To this end we followed a set of design choices when constructing our rules, with some rules having one of these characteristics and others multiple:

- **Deliberately specific wording.** Using language which, in its specificity, leaves vulnerabilities, often by omission. Since only the exact rules are enforced, anything not explicitly banned is implicitly allowed. ***Example:*** "Once the exam has started, students in the exam room must not talk to anyone else in the exam room, except the invigilators." - we both specify and repeat "in the exam room"; in this case students would be allowed to talk to someone outside the exam room including asking them to read out the answer while they write it down.

- **Deliberately vague wording.** Using language which is ambiguous in its meaning, or poorly defined in general. This gives students the opportunity to use the edge cases of a definition to their advantage. ***Example:*** "Students may not pass anything to each other during the exam." - Pass is poorly defined here; students would be allowed to show each other notes, or leave them for another to pick up later as long as they are not passed directly.
- **Omitted information.** Leaving out items of information from rules where they may be otherwise expected. If part of a rule is omitted then it is implicitly allowed by the system and students can therefore make use of it. ***Example:*** "No laptops, tablets, smartphones or smart watches. No calculators or any device that readily displays stored digits of pi. You must not use WiFi or Bluetooth wireless connections." - We've forbidden the most common consumer electronics. However by omission we allow other devices including desktops, servers, and those which support older communication standards such as radio or a wired connection.
- **Additional information.** Adding in information which is not required for the rule to make sense but will introduce an oversight for students to exploit. ***Example:*** "Students may choose their own seats. Once seated, they must stay in their seat until an invigilator gives them permission to leave. " - Here the addition of "once seated" leaves a loophole in the rule; if a student avoids sitting down, they may walk freely around the exam room.

The full ruleset is in Appendix.

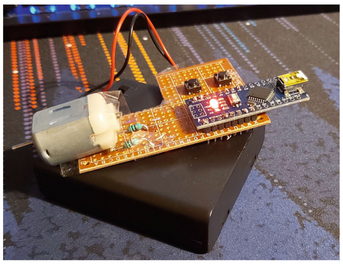


Fig. 2. A device which vibrates the digits of π when prompted. The student kept this hidden but since it doesn't display the digits it doesn't break the rules.

These rules were designed specifically to be full of oversights that students can identify, as detailed above. Also, by specifying the exam's location long in advance, students can attempt to hide things or test the feasibility of their solutions. Invigilators were aware of some of the most obvious issues with the rules, and were broadly in agreement of what constituted a rule break in advance of the

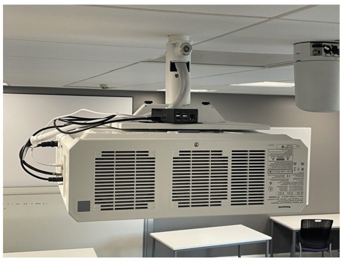


Fig. 3. A device connected to the room's projector which emulates the standard university login screen, periodically replacing the time with the digits of π one at a time. This does break the rules but wasn't detected.

exam. Finally, it's important to leave some discretion to the invigilators; easily recognisable encodings such as sixty digits of π plus one should be disallowed as invigilators see fit.

4.3 Communication with Students in the Lead up to the Exam

4.3.1 Communicating Intent

Communicating the intent of this assignment is the difference between it being a successful exercise and enjoyable experience, instead of a stressful one. Students were informed of the exam two weeks in advance during a lecture and were able to ask clarifying questions at that time, but were also referred to the module discussion page to ask further questions. Separate to the rule-set we also emphasised to students some of the intentions behind the assignment in an announcement:

> There is not one single solution to this exercise. An issue with a lot of continuous assessment is that it points you to a single correct answer but real life problems aren't like that. There are a number of solutions we expect to see from students, but we also hope that some students will surprise us with solutions we haven't thought of.

4.3.2 Answering Questions

It is important that any answers given to students were available for all students to see so that no one had an unfair advantage. We encouraged them to reach out for clarifications if they were uncertain to avoid misunderstandings.

After the announcement a lot of attack ideas were poised directly to the module team, framed as "Would I be allowed to do ...?", to which the default response was "Please see the rules". We encourage a balance between helping students and letting them navigate the rules on their own. Answers should help students gain confidence in interpreting the rules independently and reassure them that the consequences of making mistakes will not be serious. Those running the exercise should not confirm or deny any ideas, as the aim of the exercise is for students to be analysing the rule set and interpreting them themselves.

One of the initial ideas suggested was to write the literal phrase "the first 60 digits of pi". This was a valid interpretation of our initial rules and so a valid answer, but this would have been too easy and reproducible, trivialising the exercise. Students would be able to avoid creatively coming up with varied solutions. Therefore, we announced and clarified to all students that this would not be a valid solution:

> Please note however, that writing the words "the first 60 digits of pi" on the exam paper will not be accepted, we need to see the actual numbers.

We would recommend to anyone doing a similar exercise themselves to add this exception explicitly in the rules.

4.3.3 Monitoring Student Apprehension

We believe that it is important to provide students with an outlet for their ideas and questions, allowing the team running the exercise to keep an eye on them. This way if the students are heading towards an interpretation which invigilators believe defeats the intent of the exercise, it may be caught and corrected. We reinforce that part of the uniqueness of this exercise is that students will come up with unexpected solutions, and that these should be encouraged whenever possible.

As we were performing this exercise for the first time, and wanted to keep an eye on the student thoughts and sentiments around this exercise, the teaching assistants also observed the students' Discord[4] channel for the module. We found the overall feeling was that of excitement, mixed with apprehension as to possible consequences. Two questions were particularly common:

1. What happens if we get caught with a banned item?
2. Could we get reported to the academic integrity committee?

Both these questions are addressed in the initial ruleset:

> Any student seen breaking the rules by an invigilator will have to leave the exam hall and will fail on the test...
> ...If the rules outlined above are deemed to be broken, you will only fail this test, it will not be treated as a genuine case of examination irregularity.

Additionally, we used a separate announcement to brief students on how the actual day would play out as this was a common a point of discussion:

> The invigilators will act as normal exam invigilators, we will not search you on the way in, but if we see something suspicious we will investigate. When you enter the room, we will instruct you to place bags at the back of the room or under your seat.

On the Discord server, which TAs volunteered to monitor to answer basic questions, some students suggested triggering the fire alarm, along with other

[4] Discord is an online chat and call platform. There is a popular unofficial Discord server aimed at students taking the Computer Science course at the University of Birmingham, often with sub-chats for specific modules.

disruptive ideas such as using large sound systems. While we believed these to be jokes, we reminded students that national laws still apply, general decent non-disruptive behaviour on campus is always necessary, and that they should be mindful of other classes in the building. Any solutions that could fatally disrupt the entire exam experience were explicitly ruled out.

Rather surprisingly, we had some students who were unsure about whether "the first 60 digits of pi" meant the first 60 digits after the decimal point. We clarified this to the students for the sake of complete transparency. It is an option to write the desired solution in the rules for clarity, as the learning objective does not include there being a difficulty in determining this aspect of the solution in the first place.

Overall, most questions could be answered simply by directing students back to the rules. This emphasises the importance of making sure the rules are clear and represent the examiners intentions well from the beginning.

4.4 Invigilating

Before the exam, the invigilator team inspected the exam room and removed any rule violations found. Examples of actions taken include erasing writing on the whiteboard, and finding and disposing of paper sheets with 59 digits of hidden throughout the room. Students were then let in and told to find a seat, but deliberately invigilators didn't tell them to sit down. We also inspected items brought by students to make sure they did not have π written on them in the clear. To mimic a normal exam, we read out the official invigilator announcement for exams at the University of Birmingham, with minor adjustments to avoid contradicting the rules given to students.

Students ran several cheats that the invigilators could not stop, the most notable including a drone carrying a sheet displaying sixty digits of outside the room's window, and the digits of being repeatedly read out through the room's speaker system.

Students were told in the rules that if they were seen breaking the rules by an invigilator they would fail and have to leave, in actuality we decided to enforce this very leniently. Instead, we agreed to ask an offending student to stop what they were doing or put something that broke the rules elsewhere in the room. This way there was ability for them to discuss their reasoning for believing this was within the rules and if we deemed they had interpreted it incorrectly then they still could try to use other options. This meant that those with backup options could attempt them if their 'Plan A' failed. A selection of creative solutions, both photographed during the exam and afterwards submitted by students are shown throughout the paper in Figs. 2, 3, 6, 7, 8, 9, 10.

4.5 Marking and Assessment Setup

We ran this exercise as a summative assessment, but with only a token amount of credits at 2% of their overall module mark. The exam was strictly pass or fail, as students either correctly wrote all digits or not. In our exam all students who participated ended up passing. The relatively small mark value was chosen to encourage students to engage but to avoid the stress a novel and uncertain exercise may induce in students. If ran as a formative assignment we were concerned on losing out on engagement or more interesting solutions, as it "wouldn't matter". We also didn't want to harshly penalise anyone if they did fail the exercise, as the setup itself invites an inherent level of circumstance that could either help a student pass or cause them to fail. We also told students there would be prizes for:

- The most effective method
- The most imaginative method
- The best failed attempt
- The method the TAs liked best

This incentivised students to attempt more creative and sophisticated solutions in spite of potentially more straightforward strategies. We saw a wide range of solutions which we categorise and discuss in more detail in Sect. 4.7.

4.6 Educational Evaluation

After the exam, we asked the students to fill in a feedback survey, of which 68 from the 86 participating students did. Overall we had 70 responses, two of which were students that didn't actually take the exam. This quiz was optional, but we incentivised responses through making it the means by which students could submit their solution details for judging of prizes. To judge how successfully the exercise conveyed the adversarial thinking mindset, as well as whether the students found the experience enjoyable, we asked the following questions:

1. Were the learning objectives of this exercise clear? Do you understand why we ran it and how it fits in with the module?
2. Did you enjoy this exercise?

The responses for these two questions used a Likert scale, the results of which can be seen in Fig. 4. These questions were not mandatory. We had 65 responses to both of these questions, as there were some students that filled out the later questions but did not choose answers for these ones. 83 objectives of this exercise were "Very clear', and there were no recorded responses falling under "Somewhat unclear" or '"Very unclear". There were 5 responses that did not choose to specify how clear the learning objective was. In terms of the second question, the majority of students enjoyed the exam, with 83 the exercise was "Very enjoyable". There were only 7 responses that did not choose an explicitly positive response, 5 of whom didn't provide a response at all. The one "Very unenjoyable" response was given by a student that was unable to take part in the exercise as they were based in the secondary Dubai campus, which has a different module team and setup[5]. As they did not actually do the exercise, this

[5] The written response implicitly indicated disappointment that they were unable to participate in the exam, but it was also an obviously satirical comment so we refrain from removing it from our results.

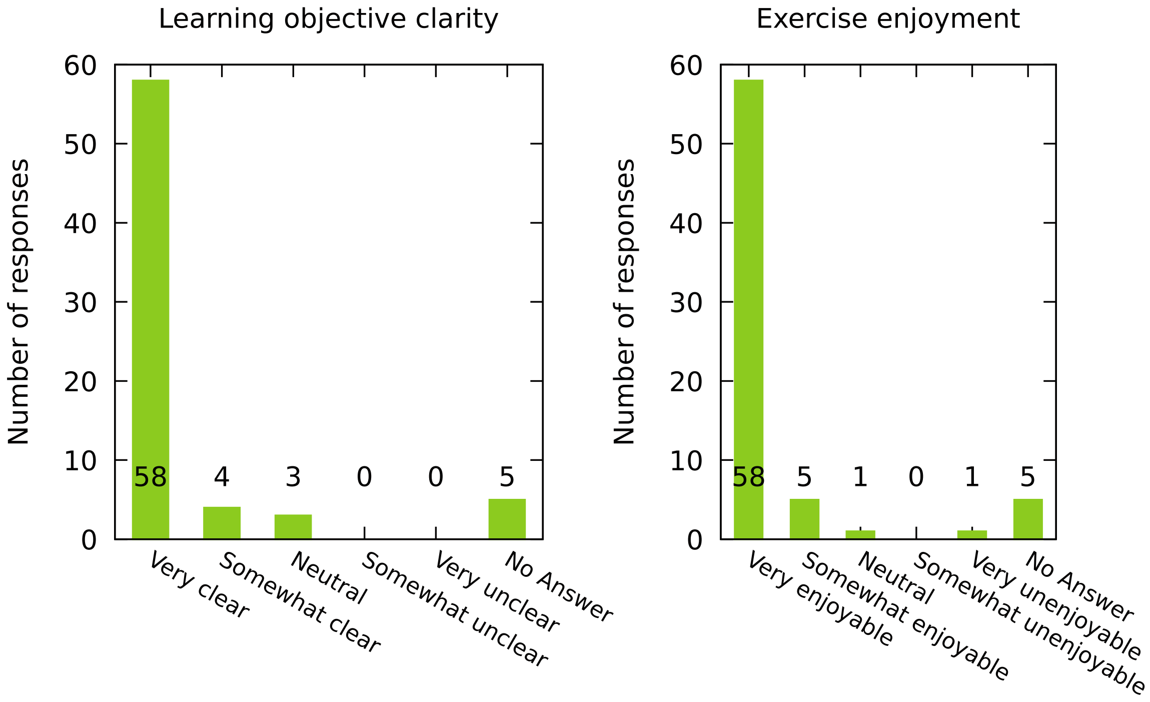


Fig. 4. Student feedback on exercise

specific response does not actually describe an opinion on the exercise. Overall, student opinion shows an obvious broad positive feeling towards the assignment, which matches enthusiasm observed on the day by the invigilators. The module team considered the results to be a resounding success on the part of the exercise.

The next set of questions concerned what solutions the student used in the exercise. These responses were freeform answers, and were not required to be filled out. We asked:

- What method did you use for your final answer?
- Did you try any methods that failed?
- Did you have any backup methods that you didn't end up using? If sop lease tell us what they were.

We used the responses to these questions to create a list of solutions. We found that many solutions followed similar approaches; these have been categorised in Table 1. These are discussed in more depth in the following section. As for the other two questions, 10 students gave meaningful answers to question 2. 7 of these answers described how a solution that they brought into the exercise failed, mostly due to the invigilators' actions. One solution (seen in Figs. 2 and in 7) was inadvertently thwarted by the invigilators turning off the projector screen as part of setup - not realising that someone had modified the display output as an attack. The remaining three answers described ideas that they had considered for their solution, but were then abandoned in favor for another solution. 36 students told us about the backup solutions that they had prepared. One student described as many as 3 backup solutions. Some of the backups described were also just alternative ideas that students had, without having actually prepared it for the day. Broadly, there was a great variety of ideas. We were happy to

Table 1. Categorisation of π exam solutions: survey questions answered by 64 students out of 68 respondents

Class	Final solution		Backup solution(s)	
Loophole	48	75.0%	32	50.0%
Encoding	24	38.5%	11	17.2%
Implicitly allowed device	11	17.2%	13	20.3%
Alternative rendering	6	9.4%	4	6.3%
Memorisation	4	6.3%	4	6.3%
Social	16	25.0%	13	20.3%
Opportunist	9	14.1%	6	9.4%
Teamwork	7	10.9%	7	10.9%
Cheat	8	12.5%	8	12.5%
Covert	5	7.8%	6	9.4%
Hidden in plain sight	3	4.7%	2	3.1%

see the level of engagement students had not only on the day, but also with brainstorming and prior preparation evidenced by the many backups described.

4.7 Categorisation of Student Solutions

As part of our evaluation process, we categorised the solutions described by students. The solutions can be broadly split into three categories. Some solutions fall under multiple classifications, as they meet the criteria for more than one class, though we generally tried to avoid this when possible.

Three quarters of the reported successful solutions methods fell under the loophole category. These were solutions that fit within the constraints of the rules when followed exactly. Within this category were solutions that encoded in an obfuscated way, so that they did not violate the rule regarding having written material (see Figs. 5,6 as examples). Other methods included in this class were devices not explicitly excluded by the rules. Smartphones were not allowed, so instead attacks made use of devices such as E-ink displays, Raspberry Pi Picos, Arduinos, MP3 players, and walkie talkies (which communicate over radio rather than Bluetooth or Wi-Fi). One high effort solution involved bringing in a server (see Fig. 8) to connect to a display in the examination room. Solutions classed under alternative rendering exploit the strict semantic reading of the rule about presentations of that were not ‘written’ or ‘printed’. Examples of this encountered was being: carved into a piece of card, baked into a pie, sewed into a jacked sleeve and painted as a mosaic. A small amount of students also memorised all required digits of π. Some had already had some amount of digits memorised before the assignment was issued, and some memorised it specifically for this assignment.

Some solutions relied on the actions of other students. In this category were people that planned ahead of time to work together for their solution, such as a group that coordinated a drone flight outside with a poster hanging from it. A solution that could have only worked with a group, but also was classed

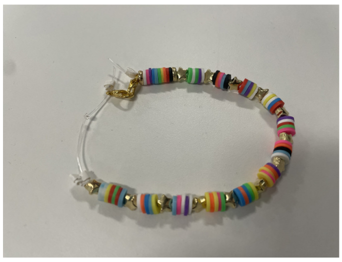


Fig. 5. A simple encoding using colours on a bracelet divided into sets of five. The first ten colours are the key used in decoding.

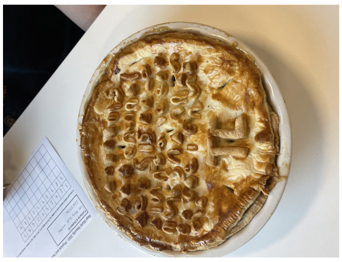


Fig. 6. A more complex encoding baked onto a pie. In this case the key is the number of edges each shape has. For obvious reasons this was a favorite amongst TAs.

under memorisation, were six students that assigned each other 10 digits of π to memorise, who then collaborated together in the exam. There were also the students that made the decision to go in without any solution, believing that the other students' solutions would provide an opportunity for them. This is a high-risk strategy, but in many cases this worked very well for these opportunists.

Some solutions we classify as *cheats*: these would be any solution that violates the rules set, however if students could get away with it they were awarded marks. If these solutions were discovered by invigilators, the student would have failed. These make up the smallest category, but were impressive as attacks because of how well they worked. The success of these solutions were based on how well they could go undetected. A small amount of these were "hidden in plain sight", using explicitly allowed items with slight modifications to hide π completely in the clear. The Figs. 9, 10 show examples of such solutions. The other cheats relied on being hidden and undetectable. One such example is where a student had written in a transparent pen in a way that the text could be hidden by their hand, and also having a second pen as 'a distraction'. Another example utilised an Arduino device (see Fig. 2) - technically allowed as a device, but the student makes a point of mentioning how they tuned it to be 'tactile but completely silent'[6].

When discussing potential solutions as part of the planning process, encodings was not something we considered as an obvious solution, so the end result where the majority of solutions incorporate some manner of encoding was novel.

[6] "... I could have it in my hoodie pocket and run it completely unnoticed, so even though it does technically fit in the rules I was guaranteed to be good. However this discreetness wasnt needed with how obvious everyone else was."

It appears to have been an approachable, simplistic way of enciphering the information just enough so as to not break any rules, but still allowing for a good amount of creative expression and variety in student attacks. Thinking about the learning objective of this specific exercise, we ideally wanted to see more loophole style attacks than cheat style attacks. These loophole style attacks can be related to e.g. traditional protocol attacks, where oversights in design allow for exploits. Such protocol attacks and other similar cybersecurity attacks feature heavily in the module this exercise was run for. That the majority of students used a loophole as their main solution indicates good understanding of the learning objective, marking the exam as a successful teaching method.

4.8 Effect on Prospective Students

In the run-up and the aftermath of running this exercise, the excitement around the module spread outside our module through to other students in the Computer Science course. In order to explore this effect, and whether it made students eager to choose the module, we sought out prospective students in the process of deciding which modules to take in their final year.

We used Discord to gauge the opinion of prospective students - these are students that would be more likely to be talking to those from other years. One of the TAs asked questions in an open forum (an open Discord channel), and received six distinct responses from students considering enrolling in the course. While we had done nothing to inform these students of the existence of this exercise, 3 of the 6 stated that they had already heard of it from speaking to students in the year above. They stated that it had made them more likely to take the course. We found that many students were already aware of this from speaking to students in the year above. There was a generally positive sentiment among these students, who were already excited by the idea of participating in this assignment. We did not receive any negative comments, or comments that indicated that they are dissuaded from taking this module due to this assignment. One student said that they "think it sounds great and was part of the reason [they] picked [the module]", another that "it made [them] more likely to

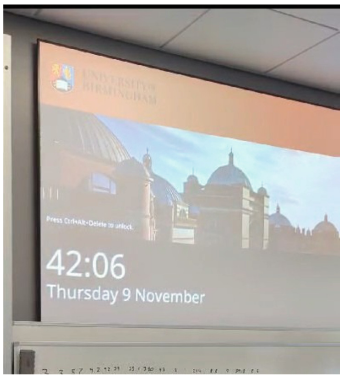


Fig. 7. Student submission: "TLDR: I created a fake login screen that would display the digits of in the time every 5 minutes after 11:20am." Shut down by the invigilators turning the projector off for unrelated reasons.

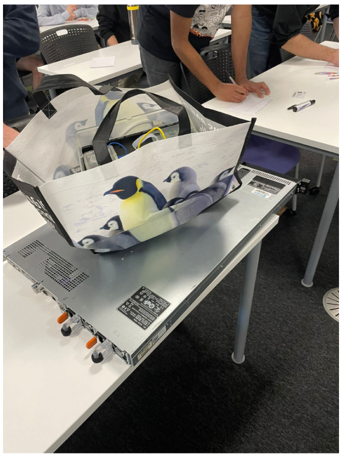


Fig. 8. A server along with the required cables and peripherals to use. Nothing in the rules forbids such a device from being used.

pick [the module]". One student said that "it's unlike any other exam I've heard of", " ...here's this set of constraints, now effectively cheat without breaking them ... it just sounds really fun". These comments then generated further discussion for those interested in this module and what their own solutions would be. These responses are reassuring that students would still engage well with the upcoming assignment, and serve as a promising start to in our efforts to gauge if the exam has a significant effect on module takeup.

These results provide promising indicators that our design has academic merit. and that it would be worthwhile to run the exercise again in some form. The module Security of Real World Systems received high praise this year on the National Student Survey (NSS), being described as engaging, well run, as well as fun and challenging. While we cannot say which aspects of the module students are referring to here, we believe this exercise had a large positive impact on the module. With future iterations we would like to record a more exact measure of the academic benefits this exercise gives students.

5 Discussion

5.1 Exam Question

In running our first iteration of this exercise, we chose to have students recite the first 60 digits of π; this was derived from Conti and Caroland's exercise in which students were required to recite the first 100 digits of π. We chose to shorten the required answer slightly in order to reduce what we saw as an unnecessary complication - we felt students that were able to achieve the required objective of 60 digits of π would have already demonstrated a workable solution and any more would be unnecessary additional work. Additionally, we considered this above the lower bound under which students would consider memorising the required answer rather than finding a solution that better fits the intended learning outcomes.

5.1.1 Memorisation

We feel this worked well, but there is room for improvement. The most obvious concern is the ability for students to memorise the question answer - in our case, this represented a small contingent of students. One student remarked that they had already memorised well over 60 digits of π before knowing the premise of the exam. Additionally, one of the prospective students interested in taking the module also has many digits memorised: "i know like 200 rn i think i could get to 1000 if i tried". We feel that this discourages the use of written prose for answering the question; we believe that prose would be more straightforward to commit to memory.

Of particular note is the use of π for the question. There is something of an understanding that π is a numerical constant the people memorise; in some circles the ability to recite it is a mix of challenge and point of pride. While a very small number of students were already able to recite some or all of the required 60 digits, it is unclear to us if this phenomenon caused an increase in the number of students that more readily reached for memorisation as a solution.

5.1.2 Repeatability

There is some thought to be put toward whether re-running the same question would be beneficial: in our discussions with prospective students they have already started crafting their own solutions. While this is heartening to see it may not be the desired outcome. There is also the possibility of students using previously successful solutions again. It should be first established whether this should be discouraged, or whether it actually is in the spirit of the learning objective (in the same way opportunism was a valid solution for various students). Generally, we want students to come up with novel solutions in order to understand the adversarial thinking mindset, so variety in the question given may prevent these potential issues.

5.1.3 Encoding

Another consideration is the ease of encoding and decoding the answer; in this case, encoding a series of numerals is fairly straightforward. Using a string that is more difficult to encode straightforwardly - consider a string with a wider range of characters in use - would discourage encoding solutions or demand that they be more sophisticated.

5.1.4 Novel Questions

Core to this premise is that students are told the question far in advance; however this too can be modified to adjust the difficulty of the exercise. Consider also a similar exercise where the question is not revealed beforehand. This demands solutions be both online and bidirectional - offline solutions such as encodings are completely eliminated and any solutions must also allow the student to communicate the question as well as receive the answer. This would raise the difficulty substantially. Another question in this vein with the same premise could be to have an established text or source of knowledge (e.g. a novel) that could not be present in the exam venue, and in the exam have the previously unrevealed question ask for a specific subsection of that text (e.g. "Give me the 4th paragraph from Chapter 7"). This question would allow for offline solutions like the question does, but having effective and easily enciphered offline solutions would be significantly more challenging. Conversely, this style of question could encourage attacks requiring teamwork, where the goal is to then communicate to the student in the exam venue the correct answer. It is worth considering how much the question itself affects the solutions that are used to attack it.

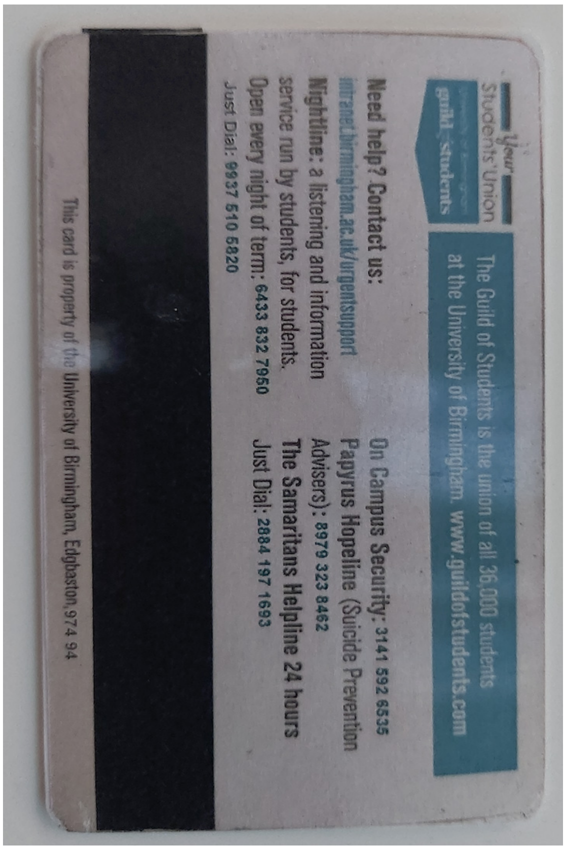
5.1.5 Configurability

The main concern of the question asked of students is that it serves to set the difficulty of the exercise and in what dimensions it is difficult. Consider if encodings should be encouraged or discouraged; also consider if multiple questions could be asked to define a moving level of difficulty. For example, by asking one question known in advance and one novel question, it is possible for solutions to be rewarded differently based on their complexity.

5.2 Opportunistic Student Behaviour

While it did not represent a significant proportion of the solutions, about 14% of solutions were 'opportunistic', which is to say that these students instead relied on eavesdropping a solution from another student.

Some of these were cases where the student did not originally intend to do so, but made do when their initial solution failed; others were students that



Your Students' Union
guild of students

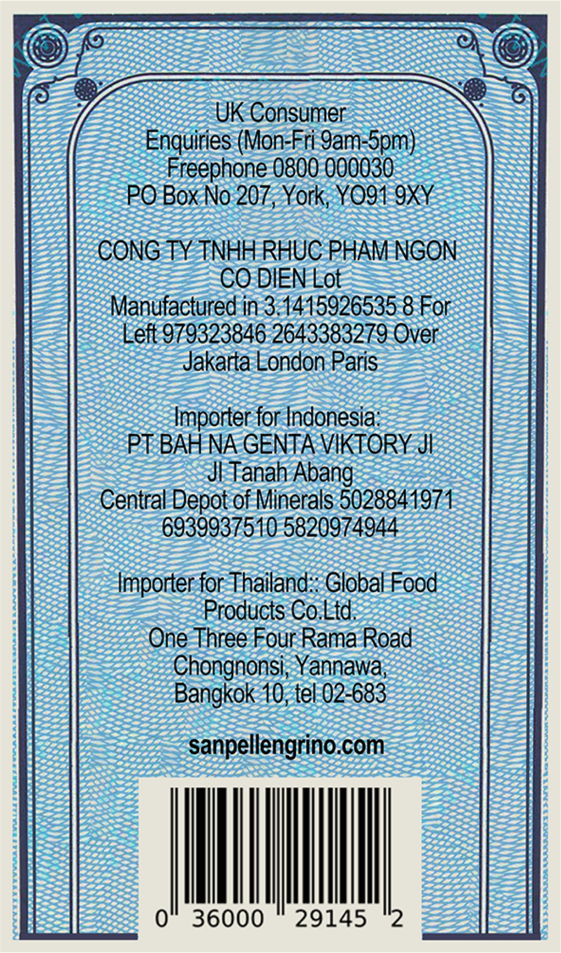
The Guild of Students is the union of all 36,000 students at the University of Birmingham. www.guildofstudents.com

Need help? Contact us:
intranet.birmingham.ac.uk/urgentsupport
Nightline: a listening and information service run by students, for students.
Open every night of term: 6433 832 7950
Just Dial: 9937 510 5820

On Campus Security: 3141 592 6535
Papyrus Hopeline (Suicide Prevention Advisers): 8979 323 8462
The Samaritans Helpline 24 hours
Just Dial: 2884 197 1693

This card is property of the University of Birmingham, Edgbaston,974 94

Fig. 9. Here the student has printed out a fake back to their ID card, replacing all the numbers with the digits of π. This breaks the rules but went undetected.



UK Consumer
Enquiries (Mon-Fri 9am-5pm)
Freephone 0800 000030
PO Box No 207, York, YO91 9XY

CONG TY TNHH RHUC PHAM NGON
CO DIEN Lot
Manufactured in 3.1415926535 8 For
Left 979323846 2643383279 Over
Jakarta London Paris

Importer for Indonesia:
PT BAH NA GENTA VIKTORY JI
JI Tanah Abang
Central Depot of Minerals 5028841971
6939937510 5820974944

Importer for Thailand:: Global Food
Products Co.Ltd.
One Three Four Rama Road
Chongnonsi, Yannawa,
Bangkok 10, tel 02-683

sanpellengrino.com

0 36000 29145 2

Fig. 10. Another hidden solution, the student has replaced parts of the text on this drink label with the digits of π. Again, this breaks the rules but remained undetected.

entered with no plan except that they would be able to leverage another student's solution.

It is unclear how desirable this behaviour is; ideally all students would have to engage in the act of examining the rules and creating solutions. However, this does appeal to the naturally social aspect of the exercise - which we feel is important - and it is unclear exactly how to discourage this behaviour without also impeding teamwork-based solutions. In running the exercise we found it an acceptable solution; we considered it a high-risk strategy. Important to note is that this is partially modulated by the chosen rules - we chose to allow students to walk around (should they notice that specific wording in the rules), which allowed more opportunistic strategies. Conversely, teamwork allows more sophisticated solutions; we feel that this should be encouraged.

5.3 Arbitrating Rule Clarification

As mentioned, it is essential to maintain an open dialogue with students between releasing the rules and the exercise day. This channel should be accessible to all students - if students contact any member of the exercise team individually, any clarifications beyond being told to re-examine the given rules should be announced to all students.

These clarifications should also be centralised so that they can be referenced by members of the exercise team before responses are given; in our case confirming every "ruling" or clarification between the team was infeasible, so every member of the team had the authority to make these clarifications.

To emphasise the importance of centralising the results of these clarifications is highlighted well by a case we encountered. We were not keeping track of these clarifications strictly enough, and one member of the team - also unable to attend to invigilate on the day - told a student that a specific device, a game console, would not be in breach of the rules. However, on the day, this was determined in breach of the rules as it could communicate over WiFi; this unfortunately meant that the student was advised in one way but the rules were enforced in a conflicting way.

While this was the only case of this we received in feedback, it demonstrates the importance of clearly and consistently arbitrating rule clarifications.

5.4 Student Incentives

In our iteration of this exercise we used two incentives.

The first of these was an academic incentive; we made the exercise contribute a small token amount to the overall module grade for students. We did this to increase engagement as we believed that the students would be more likely to participate in a summative exercise, despite its relatively small contribution to their grade; although this it's actual effectiveness in this role is not clear. The downside to making this assessment summative is the potential for student discontent; the exercise is unlike many they may have undertaken before, and perceived unfair rule enforcement - for example - is more likely to cause upset. We see no reason that this exercise cannot be ran formatively, especially in scenarios where very high student engagement is common and expected.

The second was a set of small prizes, described above. These served the secondary purpose of encouraging students to engage creatively, and for this purpose we feel that they worked well. They also served to incentivise students to describe their methods to us, which is the only means by which they could be entered to win such a prize. Anecdotally, we found that students were excited to share their solutions with us; many of those that employed sophisticated solutions were proud of them and happy to demonstrate them in exchange for the potential of winning a prize.

6 Future Work

While we believe the exercise was a success, we have also identified a number of possible future directions. First, is re-running the exercise with a different question, as discussed in Sect. 5.1, to encourage a wider variety of solutions. Most notably, we would increase the difficulty of encoding the answer. Also, we have considered re-framing part of the objective by introducing a second, unseen question, possibly with a higher mark available. This way students will have to more thoroughly break the system in order to complete both challenges. This second question would also not be achievable via an encoding, encouraging more novel and complex solutions.

In addition to alterations we may make to the exercise itself, we are also interested in further educational evaluation. Notably, more substantial data collection and analysis, including a comparison of solution complexity and overall module attainment. A key purpose of the exam was to get students to think creatively and increase engagement, therefore we are interested in measuring the effect this exercise has on engagement with regular course content in addition to student attainment.

7 Conclusion

In this paper we have described our live security exercise in adversarial thinking and creating problem solving, run at the University of Birmingham, UK. We have outlined the attacker model given to students, as well as the rules developed to encourage the adoption of complex and novel solutions. We have analysed student feedback, finding that most students both clearly understood the learning objectives, and enjoyed the exercise. In addition to analysing student feedback, we have also categorised solutions and outlined our thoughts on each. Finally, we have discussed some of the key lessons learned from running this exercise, with the hope that they will help advise other educators on how they might best deploy this exercise.

Appendix

Exam Rules

The exam rules we gave to students are as follows:

There will be one question on the exam, which is: write out the first 60 digits of pi.

The exam will be location on date. It will start at start time and end at end time i.e., a 40 min exam. No one will be allowed to enter after the first 5 mins, or leave before the last 20 mins. TAs will escort students to the rest rooms if needed (but only if there are TAs available for this).

You may not bring in anything that has been written or printed on (not including e.g., brand names on drinks, clothing, etc.). We will check the room before the exam and will remove any writing that we can identify to encode the first 60 digits of pi.

No laptops, tablets, smartphones or smart watches. No calculators or any device that readily displays stored digits of pi. You must not use Wifi or Bluetooth wireless connections.

Once the exam has started, students in the exam room must not talk to anyone else in the exam room, except the invigilators.

Students may not pass anything to each other during the exam.

Students must bring their ID card and place it face up on their desk.

Students may choose their own seats. Once seated, they must stay in their seat until an invigilator gives them permission to leave. If you need to leave, raise your hand and an invigilator will come to you.

Any student seen breaking the rules by an invigilator will have to leave the exam hall and will fail on the test.

At the end of the exam, we will collect the exam papers from the desk. There is a box to write your student ID number and name on the paper. The students whose ID and name is in that box will get full marks if the paper has pi to 60 digits on it.

You must follow all instructions from an invigilator, in particular, you must leave the exam room when told to do so by an invigilator.

Any attempt to deliberately disrupt the exam for other students is strictly forbidden.

Note: you are not actually expected to memorise the digits of pi. The key learning objective of this module has been to think like an attacker, i.e., the ability to look at a system, learn its rules and find a way to exploit the system.

The rules of this "exam" will be enforced strictly and to the letter. There are no other rules. All the normal expectations of good, honourable behaviour that are required for the other exercises on this module, and other modules, are not required here. The normal de- scription of academic misconduct and examination irregularities in the student handbook do not apply. If the rules outlined above are deemed to be broken, you will only fail this test, it will not be treated as a genuine case of an examination irregularity.

Marks will be assigned based on the exam papers. After the exam you will be invited to fill in a brief Canvas quiz to tell us how you approached this exercise. There will be prizes for: 1) The most effective method, 2) The most imaginative method, 3) the best failed attempt and 4) the method the TAs liked best.

References

1. Carlisle, M., Chiaramonte, M., Caswell, D.: Using CTFs for an undergraduate cyber education. In: 2015 USENIX Summit on Gaming, Games, and Gamification in Security Education (3GSE 15). Washington, D.C.: USENIX Association (2015). https://www.usenix.org/conference/3gse15/summit-program/presentation/carlisle
2. Chothia, T., et al.: Choose your PWN adventure: adding competition and storytelling to an introductory cybersecurity course. In: Transactions on Edutainment XV, pp. 141–172 (2019)
3. Conti, G., Caroland, J.: Embracing the Kobayashi Maru: why you should teach your students to cheat. Secur. Priv. IEEE **9**, 48–51 (2011). https://doi.org/10.1109/MSP.2011.80
4. Davis, A., et al.: The fun and future of CTF. In: 2014 USENIX Summit on Gaming, Games, and Gamification in Security Education (3GSE 14) (2014)
5. Ensafi, R., Jacobi, M., Crandall, J.R.: Gamification in security education (3GSE 14). San Diego, CA: USENIX association (2014). https://www.usenix.org/conference/3gse14/summit-program/presentation/ensafi
6. Students who don't understand information flow should be eaten: an experience paper. In: 5th Workshop on Cyber Security Experimentation and Test (CSET 12). Bellevue, WA: USENIX Association (2012). https://www.usenix.org/conference/cset12/workshop-program/presentation/Ensafi
7. Vigna, G.:Teaching network security through live exercises. In: Security Education and Critical Infrastructures. Ed. by Cynthia Irvine and Helen Armstrong. New York, NY: Springer US, pp. 3–18 (2003). ISBN: 978-0- 387-35694-5