The Association for Software Testing (AST) is a non-profit organization dedicated to advancing the understanding of the science and practice of software testing. We were founded almost two decades ago by a group of the most influential software testers of the 21st century, and have members all around the world.

The discipline of software testing rewards careful consideration of the technical and sociological aspects of creating and using software. AST was recently asked for an opinion about online examinations for certification bodies, specifically Bar Exams.

Our conclusion is that to properly serve their technical and social purposes, online examinations must be administered in a fair and unbiased manner. They should not be difficult to undergo for people of modest means, and their administration should not create additional stress on already stressed examinees due to implementation or technology. Any gatekeeping these exams represent must be based strictly on merit.

No candidate should fail an exam other than on merit. Failure grades because of equipment barriers, power outages, the widely known unreliability of the Internet, or inherent racial and class biases in algorithms and examination methodologies are unfair to both the candidates and society as a whole.

If an examination as currently planned can't meet these requirements for ***all*** examinees, then it should not proceed. Even during COVID-19, there are methods to administer in-person exams. Decisions about how to administer exams should be made to best accommodate examiners AND examinees.

Table of Contents

Context for this Press Release

Performance and Load Testing of Online Systems

Expected Load in California Bar Exams and Arrival Patterns

A History of Failures in Online Exam Software

Expected Problems with Client Hardware and Internet

Privacy and Security Implications of Installing Invasive Proctoring Software

Recording, Uploading, and Analyzing Videos of Examinees

Ethics and Privacy Implications of Video Recording Proctoring and "AI Detection"

Facial Recognition and Bias

Fitness for Purpose

Context for this Press Release

AST was recently tagged in a Twitter thread on the administration of online Bar exams:

> *@AST_News @ExamSoft is claiming that it can run a simultaneous, single-event, remote AI proctored, 2-day exam without an issue for upwards of 40,000 people on October 5-6, 2020. What is your professional assessment?*

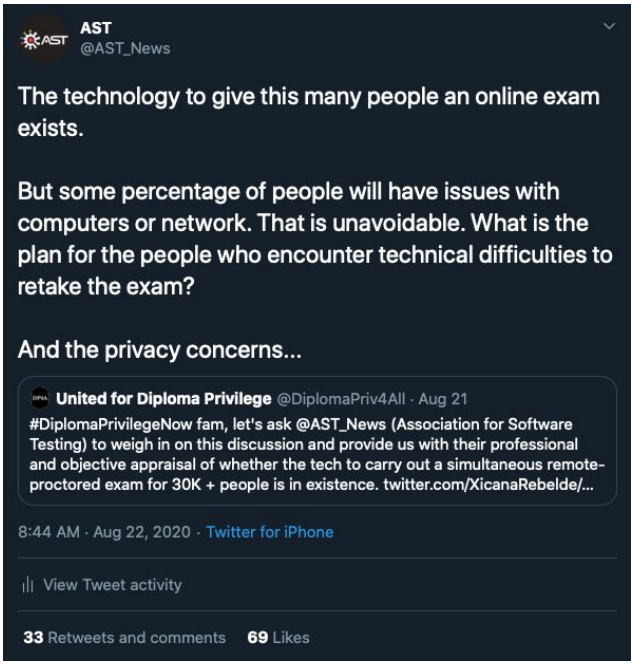
Eric Proegler, the current President of AST, replied from the @AST_News Twitter Account. He is an experienced tester who has spoken, podcasted, published, and organized peer workshops about performance testing and other aspects of software testing. The thread and additional commentary can be found later in this release.

United for Diploma Privilege, the organization that tagged AST on Twitter has since released a statement. The questions around online assessments are especially relevant to this organization for State Bar Exams, but other professions and certification bodies that use examinations to gatekeep have similar issues.

This FAQ for the California Bar Exam describes some key context around how such an assessment might be administered. The testing is scheduled over two days, with 8 exam periods that each have their own 10 minute window for password and facial recognition authentication periods.

This contract between the California State bar and Examsoft spells out a 5 year, $3.05 million agreement for examining a minimum of 12,000+ examinees annually. Some analysis of this contract can be found here.

Performance and Load Testing of Online Systems



The technology to give this many people an online exam exists.

But some percentage of people will have issues with computers or network. That is unavoidable. What is the plan for the people who encounter technical difficulties to retake the exam?

And the privacy concerns...

Performance Testing typically simulates some large number of concurrent users connecting to an online system to assess the system's ability to handle and respond in a timely fashion under the required load, before depending on the system in the real world. Canonical examples include smartphone pre-order launches, "Big Game" advertisements, Black Friday/Cyber Monday shopping seasons, concert tour and blockbuster film ticket releases, marketing announcement emails, government services websites, scheduling of courses or exams, and conducting large-scale online exams.

In cases where things have gone poorly, issues with load are often not due to the total number of concurrent users, but the rate at which they arrive. This is often called "Point Load", and occurs when many users access the same online system at the same time.

This is somewhat like attending an event in a large stadium that seats tens of thousands of people; there will be a long line outside of the gates to the stadium, and crowding in the halls as people arrive. If everyone showed up early, sat down promptly, and stayed seated throughout the game, the number of seats could be the critical measurement.

The number and locations of available gates and the reliability/speed of ticket readers, the quantity and distribution of concession stands and restroom facilities, and the width of hallways and number/location of escalators/stairs are more useful measurements of the venue's capacity limits. These will be encountered based on the crowd's arrival pattern, the length of the event, and the crowd's departure. This a more complex modelling problem than number of attendees, but also more useful for predicting how patrons of the venue might experience overcrowding.

Many online systems have worked during load testing but then failed in the real world because the load models were inaccurate. Significant accuracy in load models can be expensive and difficult to model and deliver as a Load Test. There is always some simplification, and quite often there is too much. Verifying the number of seats in the stadium is not helpful for evaluating patron experience on Dollar Beer night.

It's also worth considering how experience is measured and evaluated. Averages hide outliers, and experiences are not evenly distributed. Visitors to our stadium who have a seat near a concession stand they share with a couple hundred people from their section may be able to visit it quickly whenever they would like. Other visitors might have a longer walk to a more congested concession stand. Or they might want a particular item that is only sold in one or two places. Or the closest concession stand is understaffed, or out of the item they want. Truly understanding a system goes beyond looking at averages.

Companies like Google, Facebook, Amazon, and Netflix are admired in the technology world for their ability to build and operate online systems that support extremely large loads almost all the time. Their products still experience regular failures, despite being owned by extremely wealthy companies employing skilled engineers to create the largest, most resilient, reliable, and sophisticated online systems the world has yet seen.

Expected Load in California Bar Exams and Arrival Patterns

California's Bar is contracted for a minimum of 12,300 online examinations in two exam periods over the next five years, suggesting that the online exam system they will use must support at least 6,150 concurrent users attempting to log on and verify themselves in 8 10 minute windows over two days.

Evenly (and unnaturally) distributed, that would equate to about 11 users logging on per second. More likely is that almost all of the nominal 6,150 users will attempt to log on within 60 seconds of the beginning of the login window, meaning an average more like 100 users per second with some bursts that could reach several times that rate. If these systems were load tested, hopefully these kinds of arrival rates were part of the model.

Some online system failures under load are transient or simple timeouts, and the systems are able to stabilize themselves after the period of excessive load, merely requiring resubmission of the requests. This could be quite difficult for examinees who encounter issues, as the short window they have for logging on and the stakes will mean they are likely to keep retrying.

Other failures under load can be more difficult to recover from. Sometimes, it is necessary to reset or restart some number of services and programs to restore operations, often within a few minutes once the problem is understood. The system remains vulnerable to the same load conditions in this case, though additional system resources quickly provided can often (but not always) increase the system's capacity.

A worst case is corrupted data preventing restart of the affected pieces of the system. The Northeast Blackout of 2003 is rumored to have been caused by a corrupted journaling (detailed logging) system failure that prevented alarms being issued. These corrupted data conditions are rare, but are more likely to occur under load than normal operation - and not just proportionately, as overloaded systems expose timing conditions not present under normal load.

A History of Failures in Online Exam Software

In the world of online exam systems, there have been many, many examples over the last couple of decades of systems not holding up under load. Certification bodies typically contract with online assessment companies to administer exams, as have state and local school boards. Even companies specializing in online exams for many years continue to experience capacity and reliability problems in 2020. Here are a few references to previous load difficulties with online exam systems:

- Michigan Bar Exam, July 28: It's difficult to interpret what exactly happened, but something clearly went wrong during this exam completed by just 733 examinees. The vendor has claimed to have been the target of a DDOS attack, but there do not appear to have been any updates in the month since the exam on this claim. It is alleged that the vendor publicly posted exam passwords and emailed them as plaintext during a period of downtime.

- Pearson's history of testing problems — a list
- Diploma exam system crashes; thousands of Alberta students affected

Most evidence/results of performance testing by software vendors is not made publicly available, for commercial and competitive reasons. It can sometimes be provided under NDA by a vendor, but this evidence is often tailored to answer objections in a sales cycle, not enumerate potential weaknesses and limits.

In earlier times, customers setting up purchased software systems in their own data centers would load test critical systems after they were installed to help reduce the risk of an outage by examining performance under load. Performance-related outages by definition happen at the busiest (and therefore, usually worst) times.

These days, software is typically operated by the vendor on virtual servers leased from public cloud providers, and it is much more difficult for a customer to examine expected capacity and performance. The liability (or perceived exposure to blame) of an underpowered, malfunctioning, or misarchitected online system has moved almost entirely onto the vendor, which suits many purchasers of these systems just fine.

Claims of proven readiness for whatever demand will be placed on a system are almost always made, but these claims are often proven to be incorrect. Claims made in good faith may simply be in error, but the commercial pressures of selling large software contracts are notorious for incentivizing unethical and even dishonest business practices.

Expected Problems with Client Hardware and Internet

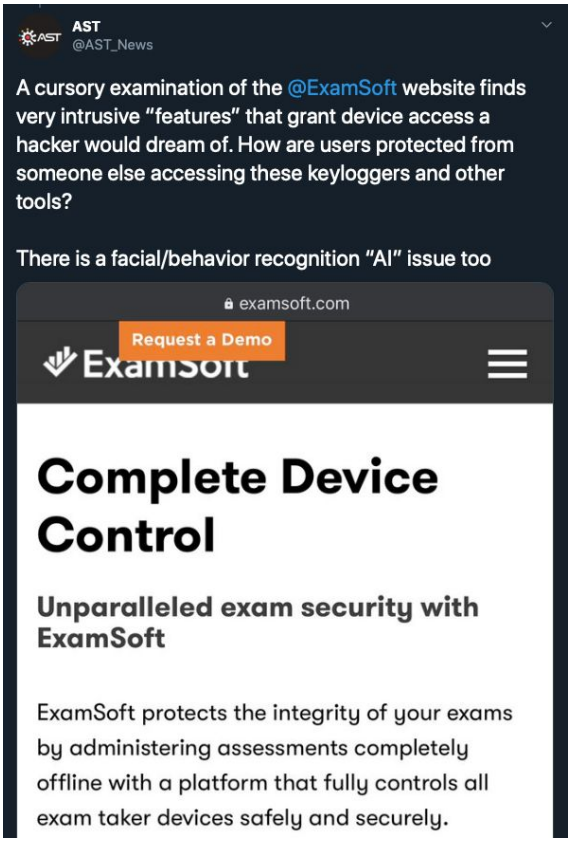
Some percentage of examinees will experience issues with their Internet connections and computers. Even the most reliable Internet connections experience occasional interruptions to connectivity, and there will be people affected by Internet outages that last longer than a few minutes. The California Bar exam FAQ insists that "At no point will applicants be able to enter the session after the close of the log-on period".

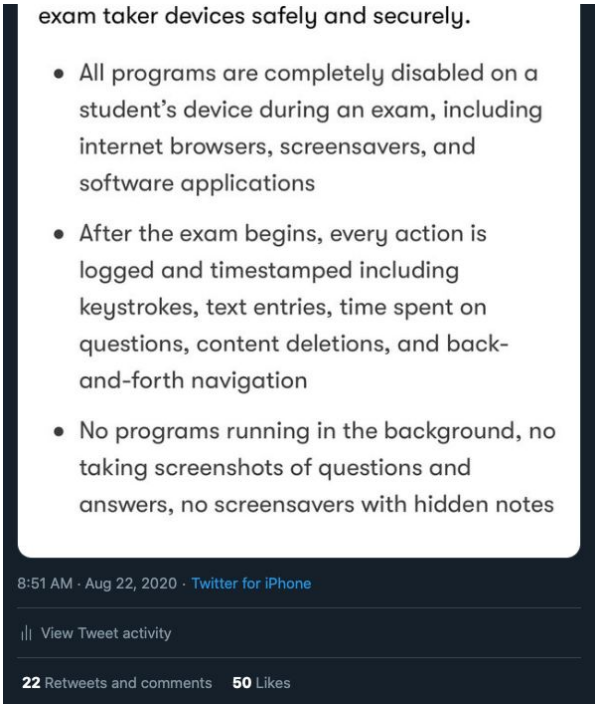
There will be people whose computers crash during the test - some of those for the very last time. Accommodation for those should be planned. How will examinees report these problems? When will they be able to retake or continue on exams?

If load issues are encountered during the test, a good outcome would be a chance to retake the test. If these load or reliability issues are not broadly experienced, reports of problems may not be properly evaluated - or believed at all.

The technical requirements for the examinees' computers can be met by most contemporary computers - meaning that examinees with sufficient means should be fine. It's not always obvious to people administering these exams (usually financially comfortable and older) what sufficient means are, and who has them.

Privacy and Security Implications of Installing Invasive Proctoring Software





A cursory examination of the @ExamSoft website finds very intrusive "features" that grant device access a hacker would dream of. How are users protected from someone else accessing these keyloggers and other tools? There is a facial/behavior recognition "AI" issue too

Software with this level of control over an examinee's computer represents a significant security risk to examinees. Installing what is essentially a rootkit makes their computer controllable by the exam software - and anyone else who can connect to their computer. While the intended use of this software is to "lock down" the computer during the test, these

tools could be used by others to steal data like passwords, social security numbers, and account numbers.

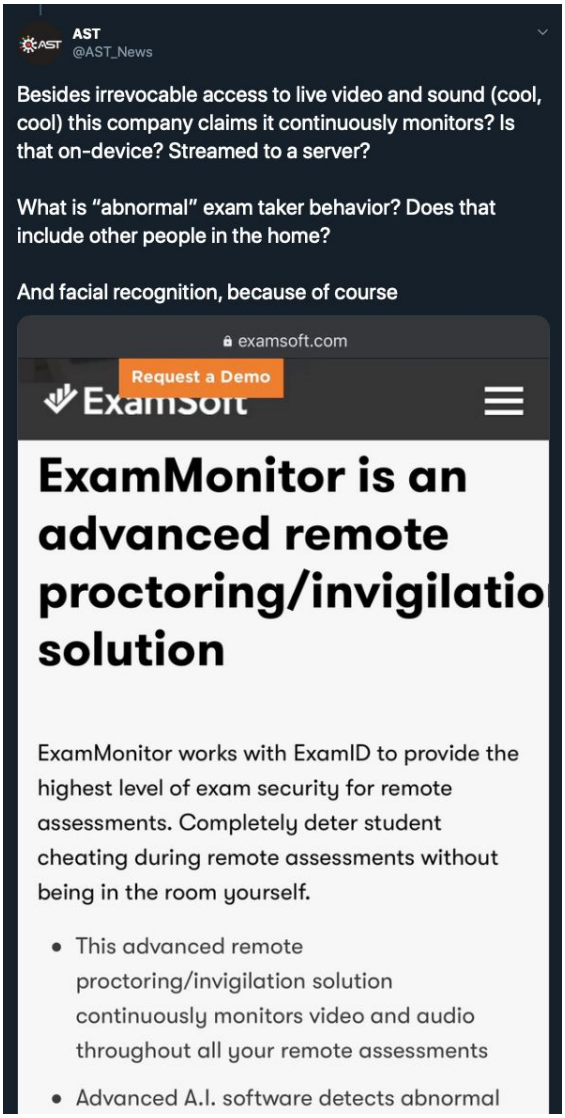
The different functions described here - blocking other applications, keystroke logging (keylogging), and shutting off wi-fi devices, are at least three different programs or components bundled together, and are probably not all written by ExamSoft. There are many versions of these kinds of tools created for hacking uses, and they are most often encountered and distributed as malware; they would be difficult to create from scratch. Copying and changing existing software typically means that the underlying code is not as well understood by the people making these alterations as if they had written all the code themselves, increasing the risk of unexpected behavior by the code.

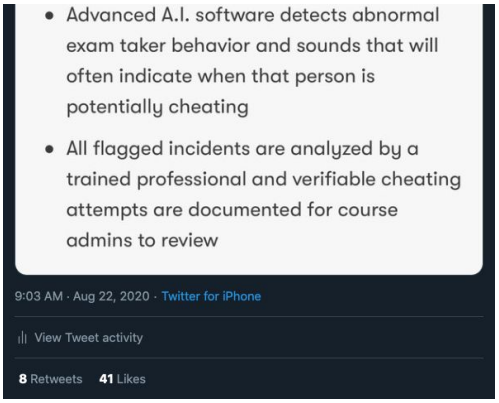
These programs achieve access to areas of the computer's operating system that the operating system is designed to protect. Examinees are required to disable anti-virus and anti-malware software to permit this software to be installed and to run, increasing their exposure to other threats while the software is present - even when not actively taking an exam.

One of the more troubling components is the keylogging component. Keylogging is a very common technique for covertly obtaining information, whether it is access to credentials for other systems, personal information, or surveillance - including cyber-stalking. It is generally accepted that installation of a keylogger is an unacceptable invasion of privacy and potentially dangerous. Two examples are people who have tried to leave an abusive partner, or are being surveilled by someone pursuing them for a relationship. "Stalkerware" can be a real threat to personal safety. Exam software installs software that exposes the examinees to that risk.

Uninstalling the exam software may remove all of these components safely. The same aspects that are designed into these tools to make them difficult to defeat and remove by the targets of the surveillance increases the difficulty and complexity in safely and completely removing them from a system. It is not uncommon to encounter issues in uninstallation with programs that are not dug deeply into protected areas of the operating system. Hopefully, this software package uninstalls correctly and cleanly, with checks to ensure all components are completely removed.

Recording, Uploading, and Analyzing Videos of Examinees





Besides irrevocable access to live video and sound (cool, cool) this company claims it continuously monitors? Is that on-device? Streamed to a server? What is "abnormal" exam taker behavior? Does that include other people in the home? And facial recognition, because of course

On further investigation, videos for this program are uploaded after the exam, and the computer's wireless connection is blocked by the exam software during the exam. "Irrevocable" is addressed in the previous section.

Examinees are generally required to upload videos taken throughout the exam period by a certain time the following day. There are technical and ethical questions about the videoing of examinees and their surroundings, and how these technologies “completely deter student cheating”.

First, let’s consider the technical challenges of capturing and uploading these video files.

Video files tend to be quite large, with size being a function primarily of resolution, and secondarily a function of encoding and compression. A full discussion of these factors is beyond our scope here, but the takeaway is that very large files must be handled. Some basic lines of questioning a tester could ask about these functions with that in mind:

1. How large are these files? How much storage space will be needed while they are captured? What ensures there is enough space on the examinee’s computer to store all the files?
2. How long will it take to upload these files? What about examinees who have slow, congested, or error-prone network connections? Are these files “chunked” into pieces that can be uploaded within a few minutes? How much bandwidth will be needed to upload these files? Are there retry and resume functions built into the uploading process?
3. What volume can the system receiving the uploads handle? At what arrival rate?
4. What operational plans are in place if the upload system has difficulty?
5. What happens if there are file errors in creating the video files? Does the program restart recording?

Different models of laptops have different built-in cameras. Part of the exam software’s proposition is that it insists the video be captured by a built-in camera, not an external camera. This seems to be intended to prevent certain cheating vectors, but it requires standards of laptop hardware that not every examinee may be able to meet. Older laptops will capture lower-resolution videos than newer ones. What about laptops that do not have built-in web cameras? According to the California Bar FAQ, their use is not permitted for taking the exam.

Ethics and Privacy Implications of Video Recording Proctoring and “AI Detection”

Video surveillance of examinees is demanding surrender of privacy rights, as a gatekeeping mechanism to future livelihood. When the demand is from a certification body like a Bar Association, there is really no choice at all. Examinees are forced to accept some number of

unknown people to look into and listen in on their homes. All of this is happening because a software vendor successfully convinced the certification body it was necessary for every examinee to “prove” they are not cheating, as a strategy to answer potential sales objections and sell software and per-exam charges. The vendor gets more data to train their algorithms, to sell even more software and assessments using this model.

This “Advanced A. I. software” is used to scan these video files and flag any evidence of potential cheating for review by a human “Test Proctor”. What constitutes that evidence is both a trade secret and information that needs to be secret for efficacy, but we can speculate on what is being “looked” for by this software. We should also consider the nature of this kind of software, and examine the claims made about its effectiveness.

Eye tracking is thought to be used to see if an examinee is looking off camera, presumably at notes or other information. People staring at a screen constantly for hours is very unnatural, so it is unclear how eye tracking would be evaluated. What if there is a nearby window they look out of while they think? What if there is a child or pet they need to watch to keep them safe and out of trouble, and probably interact with over a full day of assessment?

Facial expressions and head movements are also thought to be part of the “detection”. It is unclear how yawns, scratching, lip chewing, nervous tics or other fairly unique behaviors are evaluated.

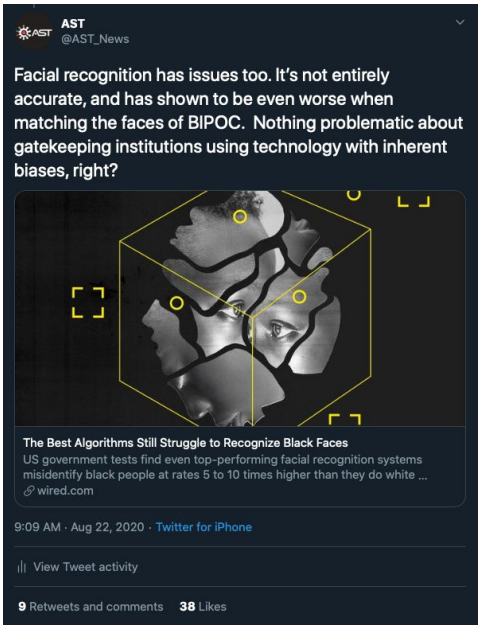
Another rumored vector is listening for background sounds as a potential marker of cheating. This would seem to be problematic - what if there are other people in the house, or a neighbor watching television or listening to music? If a person outside of a window is having a conversation with someone in a car on the curb, would that be flagged? What about a car driving by blaring a radio?

Once suspicion of cheating is recorded, the examinee becomes suspect, and the reasons why may be out of their control. Even suspicion could be risky to future livelihood; there is precedent that people believed to have cheated are barred from sitting for the exam again.

What if there is a false positive and allegation of cheating? Will that particular examinee have the patience, resources, and wherewithal to pursue appeals to an organization that has threatened their career and reputation this way? Will they even have that opportunity?

Live “invigilation” by staff of the certification body is invasive. Recordings sent to a third party are several steps more invasive than that. Using these recordings to further improve commercial surveillance software is even more invasive.

Facial Recognition and Bias



Facial recognition has issues too. It's not entirely accurate, and has shown to be even worse when matching the faces of BIPOC. Nothing problematic about gatekeeping institutions using technology with inherent biases, right?

There have been multiple recent and significant reports on facial recognition software failures, particularly when attempting to recognize non-white faces:

- "(Facial recognition) systems falsely identified African-American and Asian faces 10 times to 100 times more than Caucasian faces, the National Institute of Standards and Technology reported on Thursday." - *Many Facial-Recognition Systems Are Biased, Says U.S. Study*, New York Times, Dec 19, 2019
- "Algorithms had the highest error rates for Native Americans as well as high rates for Asian and black women." - *Why face-recognition technology has a bias problem*, Irina Ivanova, CBS News, June 12, 2020
- "The false matches were disproportionately of people of color, including six members of the Congressional Black Caucus, among them civil rights legend Rep. John Lewis (D-Ga.)"- *Amazon's Face Recognition Falsely Matched 28 Members of Congress With Mugshots*, Jacob Snow, ACLU, July 26, 2018

Within the last three months, IBM has pledged to stop working on facial recognition software. There are significant ethical issues in financially supporting development of this technology that all examination bodies - and examinees - should be grappling with.

Fitness for Purpose

Much of software testing is focused on correctness - does the software do what it is purported to, and does it behave the way its designers and developers intended? The exploration of software behavior and the comparison of that behavior to both explicit and implicit requirements is essential for safely delivering software.

Another aspect of software testing asks whether the software is fit for purpose. Does it solve the problem it is designed for? Does it serve its intended users well? Is it trustworthy enough for its intended purpose and use?

The answers to these questions can vary greatly across contexts. Software in a mobile game is probably less critical than accounting software, airplane guidance software, or a medical device's software - though revenue implications may also have a say in how much testing should be done and to which standard of quality, and perhaps even regulatory requirements if the game uses in-game currency purchased with real money.

Safety-critical software has regulatory and ethical demands of its testers. When the stakes of correct operation are life and death, the lives of each person coming into contact with the software and the consciences of everyone that worked on the software are at risk.

Financial software has a different but still high-stakes set of ethical concerns. Software errors in this sector can end companies, and ruin lives.

Software used for gatekeeping of a profession has high enough stakes in correct functionality and use to be worth examining critically. The question is, whose needs are being addressed by this software? What is the true purpose of this software?

The software maker is financially motivated to create software that sells. Reliability matters a lot to be able to sell software. Claimed features like authenticating examinees and identifying cheating help sell software. The definition of success for commercial software vendors is sales and customer retention.

The actual customer of the software could be the certification body. Their mission is typically to support a profession by attesting that aspiring new profession members have a sufficient level of knowledge and/or skill to enter it and be effective. Some of these bodies have

additional responsibilities beyond these exams, but these are usually a large part of their responsibility - and revenue.

The certification body frequently uses their position as an attestor to charge fees for taking the exam, so they profit by maintaining a monopoly on whatever is behind the gate the exam unlocks. It is in their financial interest to protect the value of the certification they provide so they may continue to sell access to it.

The certification body selects the examination vendor and writes the checks. This means the software vendor is incentivized to solve the problems the certification body has; making the exam meaningful, preventing any erosion of the examination's value, and helping the certification body manage their costs. "Solutions" like AI proctoring and facial recognition can be perceived to reduce costs when compared to human proctoring - or at least it is in the software vendor's interest to convince the certification body of this so that they can harvest some of that perceived value in software licensing fees.

Examinees may seem to be the customers of this enterprise, and they are the customers of the certification body. But they are not the customers of the exam software vendor, so their needs are not represented directly. As long as they are willing (or forced by circumstances and profession gatekeeping) to pay the exam fee, the available signal is strongly received by the certification body, and then the vendor.

Solving the problem of online exams holistically includes solving it for the needs of test-takers as well. Yes, the examination should be rigorous and its integrity should be maintained. Each person completing the exam should want its perceived value and exclusivity preserved so that it attests to their knowledge and skill - meaning they can get a job in their field and start paying off the education it took to prepare for examination.

Our conclusion is that to properly serve their technical and social purposes, online examinations must be administered in a fair and unbiased manner. They should not be difficult to undergo for people of modest means, and their administration should not create additional stress on already stressed examinees due to implementation or technology. Any gatekeeping these exams represent must be based strictly on merit.

No candidate should fail an exam other than on merit. Failure grades because of equipment barriers, power outages, the widely known unreliability of the Internet, or inherent racial and class biases in algorithms and examination methodologies are unfair to both the candidates and society as a whole.

If an examination as currently planned can't meet these requirements for ***all*** examinees, then it should not proceed. Even during COVID-19, there are methods to administer in-person exams. Decisions about how to administer exams should be made to best accommodate examiners AND examinees.

- Eric Proegler, President of the Association for Software Testing