THE (TESTING) WORLD TURNED UPSIDE DOWN

David Harley
ESET North America, USA

Email David.Harley.IC@eset.com

ABSTRACT

We often hear that anti-virus is dead, but if that is really so, where does it leave anti-malware product testing?

After decades of slow progress, security product testing has been moving away from the chaotic practices of the early 90s, to models of better practice as to some extent codified in the AMTSO 'Fundamental Principles of Testing'. Yet we've recently seen a resurgence in approaches to comparative testing that have long been flagged with a red light:

- Disabling of layers of functionality and the demotion of whole product testing
- Simulation as a comparative testing tool
- Malware creation
- Opaque sourcing, selection, classification and validation of samples
- Promotion of D-I-Y testing as superior to independent testing.

Have so many of the assumptions made on both sides of the vendor/tester divide been wrong all along? Or is this just another instance of The (Testing) World Turned Upside Down by marketing?

In this paper, we re-examine those assumptions, set in the context of:

- The good, the bad and the ugly in early product testing, and the slow-burn reaction of the security industry, culminating in the 'International Antivirus Testing Workshop' and the first steps towards the foundation of AMTSO.
- The painful evolution of AMTSO into a source of testing guidelines and, somewhat less reliably, mediation between the opposed yet interdependent testing and vendor communities.
- *VirusTotal*'s re-engineering of its policies, and the impact on AMTSO of the subsequent semi-assimilation of self-named 'next-gen' vendors into its membership.
- A new generation of conflicts between vendors and testers.
- The claimed divergence in anti-malware technologies, and mindsets, across the spectrum of mainstream and newer vendors. Does this divergence necessitate new testing methodologies? How can such methodologies appropriately be evolved, and to what extent can AMTSO successfully play a part?

Or are both AMTSO and mainstream independent testing doomed to failure and fragmentation?

INTRODUCTION

The marketing and testing of security software are two sides of the same coin. But they are not the *same* side, and their business models are not totally compatible.

There is general agreement on the need for security products, even though there are certainly some who mistrust the good faith of commercial security companies. There is also general agreement on an altruistic need to provide the general public with guidance on what products suit their environment, a need that is not purely driven by the desire to escalate sales figures. Vendors accept that testers and reviewers are a necessary part, for better or worse, of the marketing ecology. However sceptical some of us may be of the value of individual tests, it's clear that good tests (and even some bad tests) give vendors some useful feedback, whether it concerns technical issues with their software or marketing issues concerning user expectations, popularity, and perceived effectiveness in the real world. And professional testers need security products to evaluate so that they can sell their results.

Testing: the how and the who

What do we mean by 'testing'? Or to put it another way, what do we want to have tested, and where?

There are many factors that make a product suitable for a specific home or business environment, of course, but in this paper, we'll focus largely on malware detection (and, by implication, on blocking and remediation) – which isn't to say that testing of performance in terms of scanning speed and system load [1], interface comparisons, quality of support and so on aren't important, or that they aren't subject to a similar range of potential issues and problems.

We can roughly categorize testing into the three categories shown in Table 1 (see next page).

All three categories can, in principle, embrace a wide range of testing targets (detection, performance and so on) but tend to be focused on comparisons between products. Independent testing, however, can further be categorized as comparative testing or certification testing. Comparative testing is intended to ascertain and/or promote the best (whatever that may mean) product(s) out of a range, usually of products from more than one vendor. Certification testing is intended, in general, to provide evidence that participating vendors/products attain an acceptable level of effectiveness. It is 'comparative' only in that products that fail certification are assumed to be less effective than those that are successful. Some testing might be described as hybrid, in that it awards certification but also provides data that can be used to compare the performance of participating products. [6]

Independent testing has been criticized on the grounds that it uses collections of known malware or 'public malware repositories' [7] that favour so-called traditional vendors participating in the cooperative collection and validation of samples.

Shared repositories of cooperatively verified samples like WildList's WildCore and AMTSO's RTTL certainly still exist. It can be argued that they have some use in certification testing [8]. In other words, if a product doesn't detect known malware, certification failure does suggest a problem with the product. It's harder to argue for their continuing use for comparative testing, as the testing industry has always been aware, and mainstream testers have generally moved away from that approach. After all, the point of comparative testing is to differentiate between products. A test restricted to malware which is already known to vendors (or a substantial majority thereof) is not going to show enormous differences. If

Internal testing	Testing by a vendor for purposes of comparing its own products to those of other vendors is normal and even desirable. Why wouldn't you want your vendor to monitor the performance (in the most general sense) of competing products in order to see how its own products might be improved? When internal testing is made public, though, it's reasonable to expect a marketing agenda, with a consequent risk of introduced bias and even deliberate deception.
Commissioned testing	Because even the most scrupulous internal test may be perceived as biased, commissioning a test from a third party is likely to be seen as more trustworthy. As may well be the case, depending on a number of factors: • Whether the third party is truly independent. A testing lab can have a close relationship with a vendor and still be considered independent, as *Virus Bulletin* (mostly [2]) is, despite its physical proximity to and shared origins with *Sophos*. There have, over the years, been 'testers' whose impartiality and independence have been much more dubious, however. • The competence of the tester. I don't think that anyone accused *ISE*'s testing for *Consumer Reports* in 2006 of being deliberately deceptive, but many pointed out the flaws in its methodology [3, 4]. • The influence of the commissioning vendor over the test's design and methodology. One testing organization commented regarding one vendor that even if its product scored well in its tests 'they "trust" only their own sponsored test, where they can dictate the methodology' [5]. While a tester can choose the degree of control a sponsor/commissioner is allowed to exercise (even to the point of declining the commission), it can't control the use that is made of the test results in the client's marketing.
Independent testing	Tests carried out by organizations that are not joined at the hip to a single vendor and constrained to tailor their tests to show a specific product line to its best advantage. Independent testers may, of course, carry out sponsored tests as described above, but a truly independent test should not be influenced by the requirements of a specific vendor. In principle, testing carried out by a potential consumer of security products in order to evaluate products in order to select the best fit for its own organization and purposes can also be described as independent, despite the difficulties of setting up *competent* testing. Unfortunately, the independence of such testing can be compromised by importing methodologies and (especially) samples from sources that aren't demonstrably independent.

Table 1: Testing types.

comparative testing were about the exclusive use of cooperatively verified lists, it would still be more accurate than using samples supplied by a single vendor and containing a high percentage of non-malicious files. But using such lists is not what most comparative tests do. Characteristically, the big independent testers go out of their way to build up their own collections of captured malware and use them for comparative testing. Samples are (or should be) verified by the tester before the test, but by participating vendors only *after* the test by way of a dispute process.

The tester's dilemma

There's nothing new about security vendors wanting to dictate testing methodology, and even whether their products should be publicly tested at all, for example with licence wording forbidding publication of test results without prior permission.

Testing: licence to bill

More recently, vendors have denied licences to certain testers. Even more recently, vendor criticism of the use of 'unauthorized' licences escalated to talk of 'software piracy' [9] and legal action [10]. Do testing labs have a right to test any product they wish, with or without authorization? I suppose they might plead some kind of journalistic 'public interest' defence to justify obtaining a licence by the backdoor. A safer course of action might be to note that vendor X declined to test and leave the audience to draw its own conclusions (and let competing vendors make PR hay) [11].

Functionality in isolation

A real problem for testers is that detection in most of the modern mainstream commercial products is multi-layered, and a test that only addresses one or two aspects of a product's detection technology cannot be wholly accurate. People who read a review expect it to be authoritative, but the sad fact is that whole product testing is difficult and expensive to implement – which isn't what people who commission tests usually want to hear. For this reason testers tend to address relatively small areas of functionality in order to keep their tests manageable and economically viable. A significant challenge is in doing so without misleading the review reader into underestimating a product's abilities by artificially disabling functionality. A test audience is entitled to expect accuracy in the tester's evaluation of the functionality and value of the product or service.

A further challenge is working with a test set that is truly representative of the threats that are most likely to affect the readers of the test reports, and testing in a way that accurately reflects the real world and the needs of the customer. Apart from sheer sample glut, there are issues like these:

- Presenting the threat in a 'natural' context (one in which it's reasonable to expect a product to detect it).
- Finding a way to test detection dynamically in the cloud without risking leakage of undetected threats to external systems.
- Correct classification and validation of threats and appropriate configuration of the software under threat.

Testing with large sample sets may actually reduce the differences between products, whereas the tester's aim is, arguably, to demonstrate a significant difference between the 'editor's choice' and the rest of the pack. We can, perhaps, all think of examples of reviews that have magnified a small difference by, for example, using percentages instead of absolute values in small sample sets, so that a product missing one detection in a test set of ten looks 10% worse than a product that detects all ten.

A more useful way of achieving superiority might be to extend the concept of 'whole-product testing' to a far wider range of product functionality than just detection.

Anything you can do…

Who can test security software (certification testing *or* comparative testing)? In principle, anyone. While the Anti-Malware Testing Standards Organization (AMTSO) has taken steps in the direction of accrediting individual tests, there is no universally recognized accreditation of individual testers or organizations. There are ISO and other standards to which it's reasonable to expect a good commercial testing organization to conform, but they're not mandatory, and I can't think of one that is specific to the testing of security software.

In fact, anyone can say: 'I know, let's charge people for testing security software' [12]. And that usually means that you have no objective assessment of their knowledge and expertise, or indeed of their affiliations and bona fides.

Look back in… resignation

Back in the day, visitors to forums like alt.comp.virus regularly asked to be given virus samples so that they could do some product testing. Some of them may even have been genuine testers rather than script kiddies.

Since testers were not required to validate their own qualifications (least of all in alt.comp.virus), the tools they used were legion, and subject to no real controls or oversight. [13]. Journalists, academics, wannabe hackers, and testers of indeterminate independence would review products based on their own 'testing' with a range of misconceived and sometimes downright deceptive techniques. An article in *Virus News International* (*VNI*) [14] details a number of such techniques. For the purposes of this paper, I'll pass over points not directly concerned with malware detection (such as timing tests), although that isn't to say that they aren't relevant to the testing (and testing-related marketing) carried out today [1, 15]. The article focuses on two mythical products (GrottyScan and WonderScan) and the ways in which the second, superior product can be made to look inferior.

Is this my best side?

Among the techniques mentioned is that of testing products with key functionalities disabled, or not testing functionality in which a favoured product doesn't show its best side. The *VNI* article refers to several instances of avoiding testing a particular component or functionality that a favoured product doesn't have (for example a TSR scanner – remember the days when nearly all scanning was passive external scanning?) or doesn't do well (for example, file repair or detection of stealth viruses).

AMTSO has long advocated 'whole product testing' as best mimicking a product's real-world performance [16]. Yet testers continue to isolate functionalities: sometimes to follow a 'sum of the parts' model, but often in order to save resources, and perhaps sometimes to show off the best features of certain individual products. In the days when static testing (and pseudo-testing using resources like *VirusTotal*) ruled the airwaves, it was necessary to emphasise that it's not possible to evaluate a modern product's detection performance accurately without allowing malware to actually execute (albeit in a virtualized or isolated environment, hopefully) [17]. However, in a multi-layered security product, the impact of the removal of other layers of protection may be critical to correct evaluation.

In 2016, one company suggested that another vendor's public demonstration, testing the effectiveness of its own product and three rival products, against '100 of the latest virus samples and 100 mutated virus samples' [18] was stacking the deck by disabling '...key (and default) protection settings [in the rival products] ... The same behavior has been reported by multiple other vendors, including the disabling of everything other than hash lookups...' [19].

AMTSO-aligned testers have moved towards 'whole-product testing' in recent years, which is exactly the direction in which testers need to go in order to evaluate products fairly, irrespective of whether marketing departments refer to them as first-generation or next-generation [11].

Independent comparative testing is often [20] carried out using default product settings. While there are circumstances under which this may introduce bias – as when a sample set is used that includes 'possibly unwanted' programs, which aren't necessarily detected *by default* by all security software – there is nevertheless an argument for using default settings in order to get as close as possible to a 'real-world' testing environment. This isn't to say, of course, that an independent tester might not disable layers of functionality in order to test a highly specific aspect of technology, or for methodological convenience. Furthermore, it's worth keeping in mind that some independent testers also run tests commissioned by vendors and other not necessarily impartial organizations. That doesn't, of course, mean that commissioned tests are, by definition, not conducted impartially by honest testers. But they tend to be conducted according to the wishes of the client, who may also have the last word on what results are presented publicly, and how.

Naturally, security companies run internal tests on their own software and that of their competitors, as indeed they should. After all, they need to know how well they compare technically with those competitors, what they can learn from competitive technologies, as part of a quality assurance process, and so on.

When such tests are performed or reported in public, though, there's obviously a marketing agenda, which isn't a problem per se: in-house tests aren't necessarily dishonest. But it takes a marketing department of almost superhuman altruism and integrity to resist the temptation to cherry-pick only those data that favour the product line being marketed. We've become accustomed to post-truth assertions of superior performance in marketing materials, where rigorous verification of claims and statistical data is far from mandatory. Presumably, they have a significant impact on sales, even when they're based on black box testing by a company that can't be regarded as impartial.

WHEN IMPARTIALITY ISN'T ENOUGH

The misuse of multi-scanner sample evaluation sites like *VirusTotal* for comparative pseudo-testing of products is a prime example [21] of a way in which the cherry-picking of functionality and reduction of multi-layering can be used misleadingly in order to generate a 'ranking' of products by perceived detection performance. The subsequent generation of marketing collateral is not the only problematic issue, of course: there are other issues, such as the assumption that multi-scanner site statistics (and indeed a vendor's own detection statistics) are always a sound basis for the estimation and guesstimation of malware prevalence and of public exposure to risk from 'undetected' malware, and that has implications for tests where samples are weighted for real-world impact.

VirusTotal could be said to 'test' a file for malicious intent by exposing it to a batch of malware detection engines. However, since it doesn't necessarily use the full range of detection technologies incorporated into the products with which it has an arrangement, it can't be said to test *products*, or to represent product effectiveness with any accuracy.

One of the more dramatic turnarounds in 2016 took place when *VirusTotal* changed its terms of engagement [22]. The new wording includes the statement that '… new scanners joining the community will need to prove a certification and/or independent reviews from security testers according to best practices of Anti-Malware Testing Standards Organization (AMTSO).'

Under protest, some next-gen vendors eventually decided that they do need to work with *VirusTotal*, which shares the samples it receives and provides an API that can be used to check files automatically. However, some of the companies self-promoted as 'next-generation' and, claiming that their technology is too advanced to test, had pushed an already open door even wider by their own attempts to compare the effectiveness of their own products and those of 'first-gen' vendors. (For example, by using malware samples in their own public demonstrations or using *VirusTotal* reports as indicators of the effectiveness of rival products.) If different generations of product can't be compared in an independent test environment, such a demonstration can't suddenly be said to be accurate when used as part of a vendor's public relations exercise.

Aggregation aggravation

Multi-scanner site abuse can be implemented in many respects without the inconvenience of having to verify or handle samples, or pay for/configure/handle products. But there are other ways of implementing 'hands-off' reviewing. One, of course, is to commission someone else to do the actual testing, as is not uncommon in magazine reviews. As not all reviewers are equipped or resourced or technically competent to conduct their own reviews, this can be a positive strategy, as long as:

- The magazine doesn't impose methodological or financial constraints that make a fair and competent test impossible.
- The reviewer is able to read and interpret the test data with reasonable accuracy.
- The reviewer is not prejudiced by other factors such as a desire to favour major advertisers.

We have seen several instances where different reviews within the same publishing group came up with different rankings from the same data. Zwienenberg and Corrons cited a somewhat similar case where two magazines from different countries used the same test. The same product came first in one review and last in the other [23].

Of course, variations in ranking are not unexpected when different review criteria are applied to the same data. We might, for instance, see data where a product scores very well on the detection of malicious samples but very poorly on false positive (FP) avoidance. In such a case, the importance given to FP generation in the review criteria could dramatically impact upon the ranking of the product and therefore the accuracy of the review. The *VNI* article [14] cites an occasion where a product made 'unusable' by its inability to avoid FPs was rated as 'the best anti-virus product on the market.' (I'm pretty sure I remember that one.)

The situation becomes more complex and error-prone where review aggregators base conclusions (in part or in total) on pre-interpreted data from multiple sources, according to criteria that may or may not work for individual readers, and where the site is part of an affiliate network and may therefore be unduly influenced by the amount of commission paid on products reviewed. Even where an aggregator site combines aggregated test results with its own in-house testing, the site's users are obliged to rely on the assumed impartiality and knowledge of testing of the reviewers. One aggregator site claimed [24] that it uses 'AMTSO certified lab test results which are the best indicator of the top antivirus software'. However, membership of AMTSO is not a certification, nor a guarantee of flawless testing, and does not constitute a blanket endorsement of individual tests. Nor does it extend to an endorsement of organizations, reviewers and software attempting to aggregate test results from AMTSO members as a form of quasi-test.

Simulation exasperation

Here's another quote from that article in *Virus News International*. [14]

> '25. Don't use viruses at all. Use simulated viruses. Assume that the simulation is perfect and that therefore all products should detect them.'

Has anything changed between 1993 and 2017? In fact, simulation in security product testing has been contentious for decades [25]. For example, a well-known open letter from the year 2000 objected to simulated viruses, courtesy of the *Rosenthal Utilities* [26]. The signatories to that letter noted that:

> 'Thus, using simulated viruses in a product review inverts the test results … because: * It rewards the product that incorrectly reports a non-virus as infected. * It penalizes a product that correctly recognizes the non-virus as not infected.'

(However, one of the signatories of that letter represents a testing organization that has subsequently reportedly conducted a sponsored test that uses 'created' or 'simulated' malware [9].)

Later, we saw more simulated viruses, courtesy of the '*Untangled*' so-called Anti-Virus Fight Club test [27], standing as representative of many other tests and reviews

where the tester used 'simulated viruses, virus fragments, or even "virus-like" programs, whatever that means'.

Bizarrely, the *Untangled* test used several instances of the EICAR test file, which despite its name [28] is not exactly a simulation and certainly not suitable for comparative testing unless the test target is along the lines of 'does this product detect the EICAR file?' [29].

There is an AMTSO guidelines document [30] that addresses the (mis)use of the EICAR file (and also refers to terminologically derived checks Cloudcar and Spycar). The conclusion says of the EICAR file:

> 'The EICAR test file has only the most limited application to and connection with testing as the term is normally understood (i.e. comparative and certification testing)…
>
> - It's intended as an installation check, not for detection testing. It doesn't tell you whether it's installed optimally, or how it detects real malware, and has no place in a test intended to evaluate detection of real malware.
> - It's useful as an installation check in that most scanners detect it, even on platforms where, as a DOS executable, it can't execute natively … However, the way in which a scanner responds is not standardized … It cannot and should not be assumed that the way in which a scanner behaves when it detects the EICAR test file is identical to the way in which it will behave when it detects real malware.'

It's sometimes suggested in the industry that non-detection of the EICAR file should 'invalidate' a product. (That's not a position I'd want to take, since, while recognition might be described as an industry standard, it's *not* mandatory.) Similar ideas also turn up from time to time in forums [31] where people talk about which is the 'best' AV program.

Perhaps the document should also explicitly have said something like 'non-detection of an installation check file is not a reliable indicator of a product's effectiveness at detecting real malware'. And as the EICAR test file turns up even today in tests, perhaps it would be helpful if the AMTSO installation checks [32] also included a note along the lines of 'Products that fail the Feature Settings Check do not conform with industry best practice. However, the checks are not suitable for or intended to be used for comparative testing. See the document [Use and Misuse of Test Files] for a longer discussion.'

In general, the EICAR file and the AMTSO feature settings checks do not simulate attacks: rather, by convention, they demonstrate that specific functionality is enabled in a security product. However, programs that are claimed to *simulate* an attack of some sort – whether it's a virus, a firewall breach, ransomware, an APT, or some other evil entity – have been around for as long as I've been associated with this industry (and longer).

But a simulation of an attack is, by definition, not a real attack. In other words, it is someone's conception of what an attack should look like, but without exhibiting any truly malicious behaviour. In general, the security industry has long preferred not to detect simulations, as they're not only technically false positives, but are generally based on assumptions about how malware and anti-malware programs work that don't hold up in practice [29]. In many cases, they're simply based on misconceptions, but at best they're based on a snapshot view of some malware, and disregard the very different technologies that may be used by different security products to detect what may be a wide range of malware. Perhaps the best reason for not supporting them is that they almost invariably mislead people who expect them to be accurate.

Companies who do choose to detect simulators are in much the same position as those companies who used to detect objects they knew to be non-malicious, garbage files, etc. that were known to be present in commonly used sample collections, because they were afraid of being penalized in poor but influential tests. In other words, to avoid competitive disadvantage, which is understandable. But by doing it in this way, they legitimize poor tests.

It's harsh to say that most simulators belong in the same category as those random text files, snippets of source code, intendeds, variants generated by inserting bits of presumed viral code into text files, randomly patched infected files, and so on – indeed, the EICAR file itself has been commandeered for somewhat similar purposes [33] – but it's hard to escape that conclusion.

Of course, the security industry – or some sectors of it – *do* support some types of tests for specific functionalities, notably those provided by AMTSO. The difference is that these do have some (limited) usefulness as an installation check. Though even *that* usefulness is, as previously indicated, compromised by openness to misinterpretation. Detection of a *qua* simulator is not a reliable indicator of anything but detection of a specific simulation. (There may be exceptions, but I can't think of one, at any rate in malware detection.) And in my view, such an inclination is actually a black mark against the product that detects it. Unless, perhaps, it's detected in such a way that it's clear that it isn't malware, in the same way that the EICAR file is sometimes detected as something like 'EICAR-testfile-not-a-virus', or using a new category analogous to the Potentially Unwanted or Potentially Unsafe categories of detection.

Real, but new malware

What about malware created or modified for the purposes of a test? The established anti-malware industry has traditionally regarded this practice with disfavour, though perhaps if it had focused less on the safety aspects and its own ethical objections, and more on the technical uncertainties that make this approach problematic, we'd have seen less of it. There's nothing new about kit-generated malware, or malware morphed into (arguably) a new variant, of course, and the pragmatic objection to it still applies: even if you disregard the other issues, you can't be sure that what you're generating is true malware. If it isn't, then it's morally wrong to penalize a product under test (or a competing product in samples supplied for an informal test [34]) because it doesn't detect non-malware as malware.

PRACTICE MAKES PERFECT

What do we consider to be 'good practice' in testing? AMTSO has generated a great deal of 'healthy scepticism' about its own validity as an organization – and that of its initiatives – over the years, but a number of highly

experienced and well-intentioned practitioners and researchers from both sides of the security vendor/security tester divide have put a great deal of effort into compiling a document [35] that outlines the fundamental principles of testing, as shown in Table 2.

AMTSO principle	Text of principle
1	Testing must not endanger the public.
2	Testing must be unbiased.
3	Testing should be reasonably open and transparent.
4	The effectiveness and performance of anti-malware products must be measured in a balanced way.
5	Testers must take reasonable care to validate whether test samples or test cases have been accurately classified as malicious, innocent or invalid.
6	Testing methodology must be consistent with the testing purpose.
7	The conclusions of a test must be based on the test results.
8	Test results should be statistically valid.
9	Vendors, testers and publishers must have an active contact point for testing related correspondence. [The phrase 'testing related' is probably meant to be 'testing-related'.]

Table 2: AMTSO fundamental principles of testing.

Figure 1: AMTSO Guidelines Documents [36].

CONCLUSION

Vendors generally acknowledge the need for testing, but that doesn't mean they're happy with the way testing is done. If they had been, there would probably have been no AMTSO. Leaving aside the multitude of totally clueless tests to focus on mainstream testing, there are some highly contentious aspects.

Pay to play

In many cases, vendors are obliged to pay to participate in tests they don't like and that offer little opportunity for feedback, let alone to request remediation of testing inaccuracies. One tester insisted that vendors pay to see full test reports, to remediate inaccuracies, and to obtain samples of malware that were alleged to have been missed. And even then, they were expected to sign a highly restrictive agreement that would essentially stop them commenting publicly on the problems with that same test.

Bontchev observed: 'My personal opinion is that the companies whose products are being tested shouldn't pay for the tests – the users of the tests should pay for the results' [11]. Well, it's hard to disagree with that as a general principle, and it would be one way of getting round the uncertainties of sponsored testing. It would be easier to accept the mainstream testing industry's claims of independence if it were less reliant financially on the industry whose products it tests. But that would be a pretty radical shift in the testing ecology. Some large companies may pay for reports, but many more probably won't. The general public won't pay directly, though they will sometimes pay indirectly for testing commissioned by magazines, consumer review organizations, and so on. What are the chances of persuading the wider community that they should be shouldering the bulk of the financial burden of running accurate product testing?

D-I-Y or Do-It-My-Way?

Cylance's Chad Skipper has also advocated 'systemic change to the pay-to-play testing industry' [7]. I wouldn't hate that, either. However, his economic alternative appears to be based on persuading potential customers to test for themselves. Since most customers are less capable of accurate testing on their own account than they probably think they are, that means encouraging them to use tools supplied by a third party. That can only work, though, if the third party or parties concerned are truly and demonstrably both independent and competent. It worries me a lot when it's suggested that testing is so easy that anyone can do it. It worries me even more when it's proposed that novice testers get not only high-level advice and documentation from a third party, but also samples. If testers (novice or not) rely on samples from a site that doesn't disclose its sources or the other financial interests of those who run it, they have to assume that the samples are valid, in the absence of a documented validation process. Unless they're able to validate the samples themselves, in which case they're less likely to need advice in the first place. And if the samples were sourced from one of the companies whose products they plan to test, or one of their business partners, their results will be hopelessly compromised.

Or do it right

Bontchev suggests that '…all anti-virus testing outfits generally fall into two categories – incompetent and incomplete. (Of course, some are both.)' [11] Which is harsh, but not altogether unfair.

I'd actually include another category. 'Misleading' or 'Subtly biased' perhaps. There does exist a wide range of tests that are not transparent in terms of their affiliations, methodologies, sample sources, and so on. If it's a choice between independent(-ish) testers and testers whose ways and means are locked up in a black box, I'll stick with the independents using what Bontchev describes as 'a generally sound testing methodology, but necessarily very limited' and continue to advocate that the wider community does the same.

The symbiotic relationship between testers and the mainstream security industry is complex, but both industries have tried hard (in AMTSO and elsewhere) to reconcile their differences in the interests of fair testing and the best outcome for the consumer [37].

Somewhat ironically, many of the issues that are upsetting so-called next-generation vendors are the same control issues that have also blighted relationships between mainstream testers and established vendors. If more so-called next-generation vendors can engage in a meaningful dialogue with those other parties in a (reasonably) neutral context – at this moment, AMTSO still seems the best bet for such a context – rather than threatening legal action and the destruction of the current testing ecology, we could all benefit. Especially the customers. (Perhaps more transparency on testing marketing models and the economic synergy between testers and vendors would benefit the consumer, though.)

AMTSO – still in with a chance?

AMTSO's public image (and indeed its very name) remains somewhat problematical. It has not, to date, set standards in a formal sense like BSI (https://www.bsigroup.com/en-GB/about-bsi/) or ISO (https://www.iso.org/home.html) [38]. It does not say who is or isn't allowed to test, and does not prescribe testing methodologies, though it does provide guidance at varying levels of technical sophistication, compiled by people with a great deal of expertise in aspects of testing. But it's the best we've got, in terms of raising the general level of testing competence.

The anti-malware industry is certainly accountable to its customers. It could be said to be accountable to the testing industry, in that the testing industry has some claim (implicit or otherwise) to represent the interests of customers. (But we shouldn't forget that the testing industry is, as Skipper has reminded us, a 'for-profit industry' [7], not some form of state-sponsored ombudsman or some sort of charitable institution.)

Nor should the testing industry be directly accountable to the anti-malware industry. But it should be accountable. Testers (and the organizations that use their results) should clearly be accountable to their audiences for the accuracy and relevance of their conclusions, and that means transparency is not optional. It allows testers to improve their testing by introducing checks on its quality. Lack of transparency compromises the validity of a test, especially if testers refuse all checks on such issues as

- how the test was conducted
- whether the published conclusions follow logically from the data [35].

When testers evade the sharing of methodological and sample information, it suggests that they don't really think their methods will stand up to scrutiny from an informed source.

Accountability still counts

Clearly, there are issues with testing that still need addressing, and dissatisfaction with such issues (pay-to-play implementations, licensing disagreements, the risk of misrepresentation of the results of sponsored tests, and so on) is not limited to relative newcomers to the anti-malware industry [39]. But the way to improve testing is not to revert to discredited approaches, or to discount all the good work that AMTSO, whatever its faults, has achieved to date.

Organizations affiliated with AMTSO are, or should be, accountable for attempting to conform to what the membership has approved as 'good practice', including [40]:

- Transparency of affiliations and methodology
- Reproducibility of results and methodology
- Statistical accuracy based on sound metrics:
  - sample set rightsizing
  - sampling techniques
  - metrication and instrumentation
  - realistic and accurate analysis
  - bias exclusion
- Ethical grounding:
  - responsible disclosure
  - declaration of interest
  - responsible sample sharing
  - duty of care (safety)
  - clarity and avoidance of misleading statements and conclusions
- Conformance to expertly formulated and agreed standards and guidelines
- Methodological validity based on:
  - comparing apples to apples rather than melons to grapes
  - consistency of test objectives with stated purpose
  - selection of appropriate test scenarios and samples sets
- Prioritization of objectivity
  - sample currency ('freshness' and in the real world)
  - validation and verification of samples and methodology.

If these values have passed their best-by date, we are in real trouble.

REFERENCES

[1] Vrabec, J.; Harley, D. Real Performance. EICAR 2010 Conference Proceedings. https://smallbluegreenblog.wordpress.com/2010/05/13/eicar-performance-testing-paper/.

[2] Brunnstein, K. Comment on Sophos' reaction to VTC test report August 2000. http://agn-www.informatik.uni-hamburg.de/vtc/EN00X8.HTM.

[3] Hawes, J. Respecting the Testing. Virus Bulletin. 2006. https://www.virusbulletin.com/virusbulletin/2006/09/respecting-testing.

[4] Harley, D. AV Testing SANS Virus Creation. Virus Bulletin. 2006. https://www.virusbulletin.com/virusbulletin/2006/10/av-testing-sans-virus-creation.

[5] Gallagher, S. Lawyers, malware, and money: The antivirus market's nasty fight over Cylance.

ArsTechnica. 2017. https://arstechnica.com/information-technology/2017/04/the-mystery-of-the-malware-that-wasnt/.

[6] Virus Bulletin: VB100 Procedures – How the VB100 testing process works. https://www.virusbulletin.com/testing/vb100/vb100-procedures/.

[7] Skipper, C. Time to Test for Yourself. Cylance. 2017. https://www.cylance.com/en_us/blog/time-to-test-for-yourself.html.

[8] Townsend, K. Cylance Battles Malware Testing Industry. Security Week. 2017. http://www.securityweek.com/cylance-battles-malware-testing-industry.

[9] Ragan, S. Cylance accuses AV-Comparatives and MRG Effitas of fraud and software piracy. CSO Online. 2017. http://www.csoonline.com/article/3167236/security/cylance-accuses-av-comparatives-and-mrg-effitas-of-fraud-and-software-piracy.html.

[10] Townsend, K. CrowdStrike Sues NSS Labs to Prevent Publication of Test Results. Security Week. 2017. http://www.securityweek.com/crowdstrike-sues-nss-labs-prevent-publication-test-results.

[11] Bontchev, V. Some thoughts on the CrowdStrike vs NSS Labs debacle. 2017. https://medium.com/@bontchev/some-thoughts-on-the-crowdstrike-vs-nss-labs-debacle-19bc15d01a2b.

[12] Harley, D.; Lee, A. Who Will Test The Testers? Virus Bulletin Conference 2008. https://www.welivesecurity.com/media_files/white-papers/Harley-Lee-VB2008.pdf.

[13] Harley, D. Antivirus Testing and AMTSO – Has anything changed? Cybercrime Forensics Education and Training Conference 2010. https://www.welivesecurity.com/media_files/white-papers/Antivirus-Testing-and-AMTSO.pdf.

[14] Virus News International, November 1993, pp.40-41, 48.

[15] AMTSO Performance Testing Guidelines. http://www.amtso.org/download/amtso-performance-testing-guidelines/.

[16] AMTSO Whole Product Testing Guidelines. http://www.amtso.org/download/amtso-whole-product-testing-guidelines/.

[17] Harley, D. Execution Context in Anti-Malware Testing. EICAR conference 2009. https://smallbluegreenblog.files.wordpress.com/2009/05/eicar-execution-context-paper.pdf.

[18] Cylance. 'The Unbelievable Tour'. https://www.cylance.com/events-on-tour.

[19] Schiappa, D. Thoughts on comparative testing. Sophos News. 2016. https://news.sophos.com/en-us/2016/06/29/thoughts-on-comparative-testing/.

[20] AV-Comparatives. Whole Product Dynamic "Real-World" Protection Test February-June 2016. https://www.av-comparatives.org/wp-content/uploads/2016/07/avc_prot_2016a_en.pdf.

[21] Harley, D.; Canto, J. Man, Myth, Malware and Multi-Scanning. Cybercrime Forensics Education and Training Conference 2011. https://www.welivesecurity.com/media_files/white-papers/cfet2011_multiscanning_paper.pdf.

[22] VirusTotal. Maintaining a Healthy Community. http://blog.virustotal.com/2016/05/maintaining-healthy-community.html.

[23] Zwienenberg, R.; Corrons, L. Anti-Malware Testing Undercover. Virus Bulletin Conference 2016. https://www.virusbulletin.com/conference/vb2016/abstracts/antimalware-testing-undercover.

[24] Fat Security. Antivirus Software Reviews – The Best Malware Protection Tools. 2017. https://fatsecurity.com/reviews/antivirus-software-reviews.

[25] Gordon, S. Are Good Virus Simulators Still A Good Idea? 1996. http://www.sciencedirect.com/science/article/pii/1353485896844047.

[26] Wells, J. et al. Open Letter. 2000. http://www.cybersoft.com/static/downloads/whitepapers/Open_Letter.pdf.

[27] Harley, D. Untangling the Wheat from the Chaff in Comparative Anti-Virus Reviews. Small Blue-Green World. 2007. https://antimalwaretesting.files.wordpress.com/2013/01/av_comparative_guide_1-3.pdf.

[28] EICAR. Anti-Malware Testfile. http://www.eicar.org/83-0-Anti-Malware-Testfile.html.

[29] Harley, D.; Myers, L.; Willems, E. Test Files and Product Evaluation: the Case for and against Malware Simulation. AVAR Conference 2010. https://www.welivesecurity.com/media_files/white-papers/AVAR-EICAR-2010.pdf.

[30] AMTSO. Use and Misuse of Test Files. http://www.amtso.org/download/amtso-use-and-misuse-of-test-files/.

[31] https://forums.lenovo.com/t5/Security-Malware/Favorite-Anti-Malware-Anti-Virus-program/m-p/3606776#M2663.

[32] AMTSO. Security Features Check. http://www.amtso.org/security-features-check.

[33] Harley, D. Pwn2kill, EICAR and AV: Scientific and Pragmatic Research. Virus Bulletin, June 2010, p.2. https://www.virusbulletin.com/uploads/pdf/magazine/2010/201006.pdf.

[34] Townsend, K. Cylance Battles Malware Testing Industry. Security Week. 2017. http://www.securityweek.com/cylance-battles-malware-testing-industry.

[35] AMTSO. Fundamental Principles of Testing. http://www.amtso.org/download/amtso-fundamental-principles-of-testing/.

[36] AMTSO Documents. http://www.amtso.org/documents/.

[37] Harley, D. When DIY Testing isn't DIY. Antimalware Testing. 2016. https://antimalwaretesting.wordpress.com/2016/12/10/when-diy-testing-isnt-diy/.

[38] Harley, D. AMTSO Not ISO. Antimalware Testing. 2010. https://antimalwaretesting.wordpress.com/2010/07/06/amtso-not-iso-standards-and-accountability/.

[39] Kaspersky, E. Benchmarking Without Weightings: Like a Burger Without a Bun. 2011. http://eugene.kaspersky.com/2011/09/30/benchmarking-without-weightings-like-a-burger-without-a-bun/.

[40] Harley, D. (Adapted from) After AMTSO – A Funny Thing Happened On The Way To The Forum. EICAR Conference 2012. https://geekpeninsula.wordpress.com/2013/06/26/eicar-paper-10-after-amtso-a-funny-thing-happened-on-the-way-to-the-forum/.