REFLECTIONS ON RATINGS

Kenneth G. Olthoff
United States National Security Agency
olthoff@earthlink.net

ABSTRACT:

Traditional computer security ratings have focused on determining the existence of a single configuration of a given system that meets the standards of a particular gradation of the rating. The author sees the use of such ratings in isolation from real world usage data leading to a situation where system security ratings bear little or no resemblance to the level of security that may be expected in actual use. Anecdotal evidence indicates that while the configuration which passed the ratings may represent a peak, the drop off from that peak down to the typical level over a wide range of configurations is often precipitously steep. It is possible that by concentrating more attention on ratings which more accurately reflect typical usage, we may see solutions developed which may rise to less lofty peaks, but which have a far higher average level of security under real usage patterns.

1 History and Present Conditions

Over the years, there have been many ratings, criteria, assurance levels, capability maturity models, etc. designed to provide insight into the security of computing systems. These schemes have had some notable differences. For example, if one were to compare and contrast the Trusted Computer System Evaluation Criteria, also known as "The Orange Book" with the International Common Criteria system that is currently being espoused, one would find noticeable differences.

The Orange Book sets down a limited number of discrete ratings, while the CC provides a language which can be used to specify an almost infinite number of possible configurations and profiles, with emphasis on all sorts of different aspect of the problem.

The Orange Book couples security functions and assurance measures, in the belief that the best mechanism in the world is useless if you have no assurance that it was adequately constructed and it is properly invoked. Similarly, high assurance is useless if it provides confidence that an inadequate mechanism has been constructed and will be invoked. The Orange Book sets out combinations of assurance and mechanism, in the desire to balance the protection provided by each.

The Common Criteria, however, decouples function and assurance, allowing one to create, if one wishes, a profile for a theoretically strong security system design with absolutely no assurance that the stated features have actually been designed as specified, or function as advertised.

The Orange Book was conceived as part of a system where a small group of U.S. Government evaluators would examine systems destined for U.S. government use, and rule on their place in the Digraph hierarchy. The CC, by contrast, is geared toward creating a commercial security industry. In the industry envisioned, anybody can create a profile. Then, any manufacturer can build a system to match that profile. Given the system and the profile, anybody who cares to pay a testing lab for their services can have an evaluation done comparing the system against the profile.

The Orange Book, and most of the rest of the "Rainbow Series" was developed with stand-alone, centralized systems in mind. While there was a "Trusted Network Interpretation", the interpretation was in some senses a permutation of the central assumptions of the Orange Book. In contrast, the Common Criteria allows one the freedom to specify a secure kangaroo if one wishes, provided one uses the proper terminology, and can adequately state the threats, goals, mechanisms, procedures and countermeasures which apply to the kangaroo security domain.

It would seem that given the marked contrast between these two approaches, one could come to several conclusions about the idea of security ratings for systems. One might theorize that there has been an evolution from the Orange Book to the Common Criteria. One might view them as opposite ends of a pendulum swing, with the current practice at any point in the future predicted to fall somewhere along the spectrum. One might view them as different business models, and debate the merits of each, perhaps using analogies and contrasts to the history of the Orange book to attempt predict the economic success and market acceptance of the Common Criteria.

In summary, the contrast between the Orange Book and the Common Criteria seems significant. One could probably find similar dissimilarities between either of these schemes and any of the other schemes that have been espoused over the years. The one detail that would probably be overlooked is that in a critical way, they are all variants of the *same scheme.*

One might ask how that could be, given all the fundamental differences in structure and philosophy noted previously, and assuming that similar fundamental differences separate all the

New Security Paradigm Workshop 9/00 Ballycotton, Co. Cork, Ireland
ACM ISBN 1-58113-260-3/01/0002

various schemes that have been proposed over the years. How is it that they all are brought together, from such scattered points on the hypothetical graph of the solutions space?

The answer is that they share a common premise. They all focus on determining whether there exists *at least one configuration* out of the vast space of all possible configurations which meets the rating, level, criteria, objective, goal, or whatever other term is used to denote demarcations between discrete categories in the particular grading system. In doing so, they ignore all the possible ways in which the system can be used incorrectly, and the various biases, assumptions and influences present in the product or system itself which may serve to guide the user to incorrect usage.

The question I feel needs to be explored is what other categories of metrics we might use. Is the existence of one pristine state of grace amidst a vast purgatory of questionable, undecidable, or just flat out bad configurations sufficient grounds for us to bandy about the word secure? Given the use of these conventional schemes to predict the existence of this one immaculate state, what sort of metrics or analysis might provide us insight into whether that state is ever likely to be achieved or even desired by an actual user community? How do we determine if it is even a usable state, in any practical sense? To use an analogy, is an impenetrable, hermetically sealed, environmentally controlled safe of any interest whatsoever if you can't fit the item you wish to protect inside?

In the interests of stimulating thought and possibly discussion, I would like to steal the basic concept of a recurring feature in the pages of MAD Magazine back in my misspent youth, and list a few "Security Ratings We'd Like To See".

2 The A. M. Best Rating

In the U.S., and perhaps elsewhere in the world, A.M Best is a firm which analyzes the financial soundness, claims payment, and general quality of insurance companies. This is ***not*** that A M Best.

The "A M Best" that is being referred to is a system espoused by a former professor of mine on the subject of user interfaces. He suggested that in order to test a user interface or other aspect of a system in something more closely resembling actual use, one needed to eliminate bias factor caused by the designers understanding how the system is supposed to work, and avoiding doing the things which would break it. To determine such "average" usage, my professor suggested "reach out in the hall, grab any moron who happens by, and sit him down to try it."

In this instance, the name A M Best refers to the idea that we want to find out what is the *Best* security that might be achieved by *Any Moron* who sets up the system. This is also phrased more diplomatically as a "naïve user test" or "analysis of typical usage" or some such.

It is common knowledge in the testing community that it is desirable to have a separate and independent test team for just this reason. The additional nuance that applies in the security realm is that in most cases, there are two aspects to the performance of the product. The first is the nominal functionality of the product, whether it be an operating system, and email package, a Web browser or whatever. The second is the security function underlying or enhancing the intended function of the system. The problem is that security products are usually designed and built by security experts, and used by those who may not understand or even care about security. The goal is to determine the extent to which the security functionality is properly brought into play by the typical user, even if they are not aware of using it.

The simple concept here is to figure out whether the security functions are used at all, whether they are used properly, and the general user awareness of the security features. The assumption is that perfect security is of little use if the users can't or don't use it. The mere presence of adequate technology is not enough, if the intended user community is incapable of understanding or applying the technology appropriately.

Note that it is not a given that the naïve user must always be aware of the security features. In some cases, it may be possible to design a system where the security functionality is effective even while remaining totally transparent to the user. This rating might also be designed to give insight into the tendency of users to turn off security functions, either for ideological or philosophical reasons, or based on the perception (accurate or not) that "security makes the system slow." The main thrust of this rating remains the same over all these possible cases – figure out how a typical, non-expert, untrained user makes use of the system.

3 The AP news feed rating

The theory is simple. For an evolving and hopefully improving field of study such as security, there are two items of interest at any given point in time. The first is the state of the art, which is the best that anyone, anywhere, knows how to do. The second is the state of the practice, which covers what the typical knowledgeable practitioner is actually doing as standard operating procedure at a given point of time. Suppose the state of the art is at the level of neurosurgery and genetically engineering solutions to disease. If the typical medical procedure remains at the level of bleeding with leeches performed by the local barber as a universal specific, the patient's health may still be at risk.

The idea would be to examine the references in the learned journals, advanced research, and proclamations from the gurus on the topic of a particular system or concept. Having determined that high-water mark as a point for comparison, one would then compare it to the articles and discourse in the popular press, the consumer level magazines, and the introductory courses at the local training centers. The ideal is for the *state of the practice* to lag the *state of the art* by as little as possible. Given that we are still reading about new iterations of security flaws first identified over thirty years ago, it is assumed that the proper scale for this metric might need to be exponential or logarithmic in nature. It is very common for the state of the practice in any field to significantly lag the state of the art, but in some circumstances, that lag is more dismaying than in others.

This rating would also lead naturally to discussions of where our energies may best be applied to advance the cause of practical security. It may be presumed that in some cases, the state of the art may already be sufficiently advanced that further research in expanding that boundary may not be the ideal course. Instead, it is entirely possible that more effort is needed in bringing the state of the art into common understanding and usage before any further expansions of the state of the art are undertaken.

4 The Arthur C. Clarke Magic rating

The renowned speculative fiction author Arthur C. Clarke is famous for his comment that any sufficiently advanced technology is functionally equivalent to magic. This rating is geared toward exploring what typical technologically advanced users are capable of doing with a product or system. In other words, if a traditional evaluation looks for at least one secure configuration, this rating model would give the product to security savvy individuals and see if they can figure out on their own how to configure the system to approach that theoretical ideal configuration. By comparison, the A.M Best model tried to find out what naïve users could do if left to figure things out on their own. This model attempts to determine if the documentation and human factors of a product or system will allow those who do, in fact, know what they are doing to achieve reasonable security using the product.

The underlying theory is that if the gurus can't get the thing configured properly, what chance does the average user have? If the experts don't understand it, then from the perspective of the average user, it is magic. We all know that there's a trick, but if you don't know how the trick is done, it's magic. The other problem with magic is that one must pay the person who knows how the trick is done to perform the trick. There is no magic without the magician, and a magician who knows tricks that nobody else can do tends to be pricier than the more run-of-the-mill magician is.

Also implied are a variety of subtexts – is the model used counter-intuitive, even to those skilled in the field? Is it practical? Is the terminology used in conventional ways, such that the gurus do not need to translate the documentation and instructions into more conventional parlance? In effect, can the "magic" be brought back down to being a usable technology, or is it irredeemably obscure, and thus of little utility in actual day to day usage?

5 The Cracker Jack rating

As many recall from their childhood, Cracker Jack is the caramel-coated popcorn treat which has as its slogan "a surprise in every box". The same might be said of the default configuration of many computer systems and software products. Unfortunately, even those systems that can be configured securely may come out of the box with default settings that are less than ideal from a security standpoint. This rating would specifically examine the default configurations of the product to determine just what surprises wait inside the box for a new user.

This is in contrast to the ratings described above, in that this rating looks at a particular state of the product, and makes no speculation as to how likely it is that either a naïve user or an expert would change the defaults. The ground rules are simple – what comes out of the box is what gets examined. While notations in the documentation about how or why to change the defaults may be taken into account or mentioned in an appendix, the thrust of this rating would be to examine the default configuration with no modification. By contrast, the previous ratings focused on the actual behavior of various categories of users, which may in fact include changing the defaults.

It may be desirable to use this and the other ratings in an iterative way. By trying the test underlying the prior ratings with various different default configurations, it may be possible to fine tune the choice of default settings to optimize the balance between security and other functionality over the broadest range of usage scenarios. This might be done by the vendor, with the aim of improving the "out of the box" performance of the product or system, or by an organization, with the intent of creating a default installation configuration for a particular environment which optimally meets the organization's policy and performance requirements.

6 The Domesday Book

This is the commonly used name for a census and audit commissioned by William the Conqueror to find out all there was to know about the new neighborhood back in 1066, the local chapter of Welcome Wagon having proven somewhat inadequate. It is included here because the idea of tracing the roots of computer security to 1066 provides the sort of reference to the history of privileged users not covered in most computer security texts, though some references to norming conquests do show up in database texts.

Anyway, imagine a Domesday Book covering a computer system – one wherein all the properties of the system, both officially authorized and commonly claimed, would be written down as they were surveyed or discovered. All contradictions between the official tallies of properties and the actual facts of properties commonly acknowledged would be disclosed, for the perusal of those attempting to administer the system. Admittedly, archives of such information do exist for some products and systems, but such archives are seldom unified and complete. William presumably had fewer problems in this area, given his position as absolute monarch – it's good to be da king!

The point remains that there is usually a great deal of information about what's actually or potentially wrong with a system. That information is of at least as much interest as the certification of a combination of factors which add up to at least one thing being right about the system. There is also the issue of parts of the system which are akin to maps circa 1066, with large blank areas labeled "there be dragons here," or worse, filled in with fanciful and wildly inaccurate documentation. In either case, the idea of this rating would be to provide as complete a documentation of the actual properties of the system as is possible.

7 The Edgar A. Poe Memorial Alarm and Interface rating

Despite the temptation to play on Mr. Poe's rather somber reputation, this rating model does not concern itself with the degree to which a product and/or its documentation resemble a horror story. Instead, this model is named in reference to a poem by Mr. Poe entitled "The Bells." For those who are not familiar with the piece, suffice it to say that the removal of the word "bells" from the poem would reduce the length of the poem by a considerable percentage. The significance of this is that many systems, particularly in the area of intrusion detection, suffer the same problem – the repeated sounding of bells (literally or figuratively) to such a degree that the operator eventually begins to ignore the bells. Having grown tired of false alarms, she starts ravin' about the system's inadequacies, and eventually says "nevermore", and turns it off.

The basic concept is to examine the manner in which a system or product indicates insecurity to the user, and the efficacy of the notification. There are possibilities for failures on two fronts. The system may not perform the underlying function of correctly identifying insecurities and suggesting remedial action. This would include inaccuracy of the information provided, or a failure to provide information to the user at all.

The second area of interest in such cases is the human factors aspect of the feedback provided to the user. Is the information provided in a form the user can understand? Does the presentation add to understanding, or hinder it? Is the interface annoying to use? It is entirely possible for the underlying security functionality to be operating properly, but to have its effectiveness undermined by a poor interface. Similarly, a wonderful interface may mask a faulty security engine underneath. Both factors need to be operating correctly for optimal usefulness.

8 The Fred and Ginger and Harold and Fayard rating

Fred Astaire and Ginger Rogers. The very names conjure up images of elegance and grace. We've all seen the movies, and the pair has become an enduring cultural icon of style on the dance floor. So what does that have to do with security?

Let us make Fred the representation of how we would hope our systems operate as a norm. Smooth, effortless motion was his trademark. He made it all look simple, and the hard work, dedication, and devotion to his craft were hidden behind a façade of spontaneous talent. We wish all our systems would make it look that easy, hiding the gory details from the users.

Ginger Rogers is famously quoted as saying that she did everything that Fred did, but did it backwards and in high heels. A true observation, that. The question is whether the security functionality we add to a system turns it from Fred to Ginger. Are we adding so much overhead that the user now has to do things backwards, and in virtual high heels? If so, is that acceptable? We should also remember that there were some things that Fred did solo on the dance floor that even Ginger could never have done backwards in heels. We need to consider what aspects of our system's normal operation are inherently changed when we add or turn on security functions.

Our ideal, then, for this rating may not be Fred and Ginger, but rather Harold and Fayard, the famous dancing Nicholas Brothers. Using them as an analogy for the before and after states of our hypothetical system gives us the desired goal – whatever Harold could do, Fayard could do right beside him, and vice versa. This is the way we want our systems to work – with the same functionality presented to the user after the addition of security as before. Granted, this will require careful design, but this sort of transparency of security and reduction of overhead is the ideal we must strive for. Rating the cost of security over a baseline is the first step in finding ways to reduce the operational impact in appropriate ways.

9 The High Tech Hippocratic rating

We are all familiar with the line from the Hippocratic Oath taken by medical doctors, which states "First, do no harm." A measurement of simple adherence to that principle might be of great benefit in the security field. Even simply pondering that one phrase on a regular basis might be a step in the right direction.

As it is, though, many systems which purport to advance security either fail to work as advertised, contain back doors which might be used against us, or otherwise interact with our computing environment in ways which cause a net decrease in security. We need to assess our systems not only for doing good, but for doing harm as well. We must then weigh the possibilities in the balance.

We are not talking about the sort of "value neutral" degradation of normal function discussed in the Fred and Ginger and Harold and Fayard rating, but rather serious harm – disclosure of data, destruction or alteration of files, or the lowering of defenses. To stretch the medical analogy, before adding a "vaccine" to our systems, we should have some idea if it will instead trigger the onset of other diseases, or even cause an epidemic.

10 The "Idiots Make Good Pilots" rating

A professor of mine who was a private pilot once launched into a discussion of piloting as a metaphor for a whole category of tasks which emphasize methodical procedure and acceptance of rules over what the participant believes they know. As Dr. Miller put it "Sometimes, idiots make better pilots, because they just do what they're taught, and don't over-think the whole thing. There are some places where being bright and thinking too hard can get you *into* trouble, instead of getting you out of it."

This same observation can sometimes be applied to security. There are some models of security that may be counter-intuitive, or otherwise deviate from the accepted norms of the field. In such cases, a practitioner with experience in the application of other models may be at a disadvantage. The prior knowledge brought to the task may be more of a hindrance than help. We have all heard of situations where individuals have been told that the first step is forgetting or unlearning paradigms learned elsewhere. In the security realm, it would be useful to identify the existence of such situations. If this type of cognitive dissonance is common when using a particular product, it would be useful to be aware of that possibility going in.

11 The Princess Bride Effective Use of Language rating

This rating covers the problem where the difficulty is not in the underlying concepts, but rather in the non-standard use of language to convey the concepts. As Mandy Patinkin's character in the film "The Princess Bride" (based on the book by William Goldman – I do not know if a similar scene appears in the book) observed, "You keep using that word. I don't think that word means what you think it means." In the ever-present search for market differentiation, some have taken the path of obfuscating or unconventional language.

While it is unclear how one might formulate a rating to measure the probability or degree of either the linguistic or the previously noted cognitive variety of such problems, it clearly would be another useful data point if a metric or rating for this phenomenon could be developed. Such a rating could be considered in the overall process of evaluating one's security options and comparing various possible combinations of products and procedures as a solution to a given problem.

12 The Ron Popiel "As seen on TV!" rating

For those who have not been exposed to the products of the amazing mind and career of Ron Popiel (founder of Ronco), let us say that Mr. Popiel is perhaps the most prolific purveyor ever known of clever gadgets which meet needs we didn't know we have. The Popiel Pocket Fisherman, the Amazing Veg-O-Matic, pasta machines, food dehydrators, counter-top rotisserie ovens, and a little gizmo to scramble an egg inside the shell are but a few of the consumer "necessities" perpetrated by this master of guerilla capitalism.

So what does this have to do with security? Everything! The security marketplace is filled with products just begging for the Ronco infomercial treatment. The breathless touting of the marvelous features, the amazing breakthrough technology, the assurance of reasonable pricing and ultimate fulfillment, and the wise, concerned, and benevolent presence of the innovative inventor are all there. And like most Ronco products, there is the tendency to overlook the basic question – why do I need this gizmo, let alone the Ginsu knives that they will toss in if I order now?

In security, as in most other areas, reasonable consumers should look beyond the wonderful demonstrations and the amazing benefits promised, because the most amazing application of technology is ultimately of little long-term value if the need met or the problem solved is not one of direct concern. We need to fight the urge to "order now!" Instead, we need to take the time to soberly reflect on how the purchase or development of the item in question will actually improve the security of one's system. Is it the answer to my real problem, or is it a flashy answer to a problem that I don't really have?

In short, we need a rating which guides us to the best match for our actual needs, rather than the best match to a hypothetical set of needs concocted by the vendor, or a consultant, or some other source of reputed wisdom. While the Common Criteria may prove to be a step in this direction, it retains the focus on "at least one correct configuration". What is needed is a broader view of the appropriateness of the product or system over a wide range of configurations.

13 CONCLUSION

Despite the fanciful names, it is my hope that the previous descriptions have brought out at least some of the aspects of the total security product usage picture that are not adequately described using the ratings and metrics which have been historically favored. We are still lacking in useful ways to determine the likely (or realistically achievable) level of security that the system may provide under actual use in typical environments, given users with typical ranges of skill in the operation of such systems.

This is not to discount the usefulness of the traditional metrics and ratings that focus on the existence of a single secure configuration. Rather, I wish to point out that while they may be necessary, they are certainly not sufficient.

Even if something like the ratings described were available, though, there are limitations that bound our possible success. Unlike many other products, security tends to be very dependent on the specifics of individual cases and environments. Security is also very dependent on the skill and behavior of the users. Combining the two, we run into a situation where it is very difficult to gather data on all the possible interactions between system, environment, and user. Once one gets beyond the simplest of security mechanisms, the possible permutations multiply too rapidly – one can gather data, but it is difficult to generalize from that data.

This is a particular weakness when we attempt to apply quality improvement techniques developed in the high-repetition, low complexity world of the production line to the long lead time, unique case world of complex system development. Statistical sample size looms large as a problem in any attempt to quantify security as we are currently addressing it. We may need modeling tools and metaphors more aptly fitted to the nature of our problem.

Additionally, many of the rating factors noted above are inherently imprecise, which leads to a sort of "Princess Bride Effective Use of Language" problem for the ratings themselves. Not only do we need ratings that tell us more about the actual effectiveness of our systems; the ratings themselves must be understandable.

We've been told "the decision maker wants a number." The problem is that if we give the customer a number, or a simple grade, the customer may not know what the units of measure are, whether the scale is linear or logarithmic, the range of possible values, or how accurately the rating is calibrated. Of even more importance, the customer may have little grasp of what rating equates to "good enough" for the particular situation.

Whatever the rating is, we need to make sure that its meaning and limitations are clear, lest our ratings take the form of pseudo-random numbers expressed to five decimal places to give the sheen of scientific precision to a blatantly imprecise measurement. The goal must be a rating which guides the decision-maker to an appropriate, informed decision.

Lastly, we must develop a firmer focus on the correlation between our ratings and end results. The current state of the art in ratings and evaluation of computer security is focused on the inputs. We analyze the development process, adherence to standards, and other evidence relating to how the thing was designed and implemented. We have done precious little work, however, to correlate the perceived benefits of those input modifications to the end results. We do not have good data to tell us what factors on the front end have the most influence on the ultimate secure usage of the system.

Even when the finished product is considered, we are at best looking at defect rates and other measures related to the quality of the development process, not to the security of the system in use. Were we designing racing cars, we would be in effect pontificating on all aspects of the design of the racecars without ever looking at the race results achieved by the fruit of our labor. We must establish that linkage between our processes and standards on the one hand, and the outcomes in "real world" usage on the other.

I do not claim to have the answers to the questions I pose. Perhaps by starting the discussion of how one might appropriately address some of these aspects of the problem, though, others may be inspired to identify the most potentially useful paths for future research. I also hope to inspire those in the security field to put forth proposals of how we might fill in these gaps in our understanding of the total usefulness of available products, systems, and procedures.