北约网络安全防御演习（Locked Shields）的评分机制

此前已经介绍过可能是地表最强网络安全防御演习的“Locked Shields”，感兴趣的请移步过去的文章，此处不再赘述。许多网络安全防御演习，特别是国际间合作的网络安全防御演习很多都反对评分制度，毕竟得分不是网络安全防御演习的主要目的。例如 ENISA 组织的 Cyber Europe 就不使用评分制度，但 CCDCOE 组织的 Locked Shields 则坚持使用评分制度。

通常网络安全防御演习都被设计成游戏化的竞赛，以提高各方的参与度。合理的评分机制可以提供有价值的反馈，并且维持参与者的热情。不合理的评分机制可能会引发各方的不满，分散参与者对主要目标的注意力。但实际上，网络安全防御演习的规模越大，设计合理的评分机制也就越具有挑战性。**衡量评分机制是否合理，通常可以参考四个维度：复杂性、透明度、竞争性与自动化。**

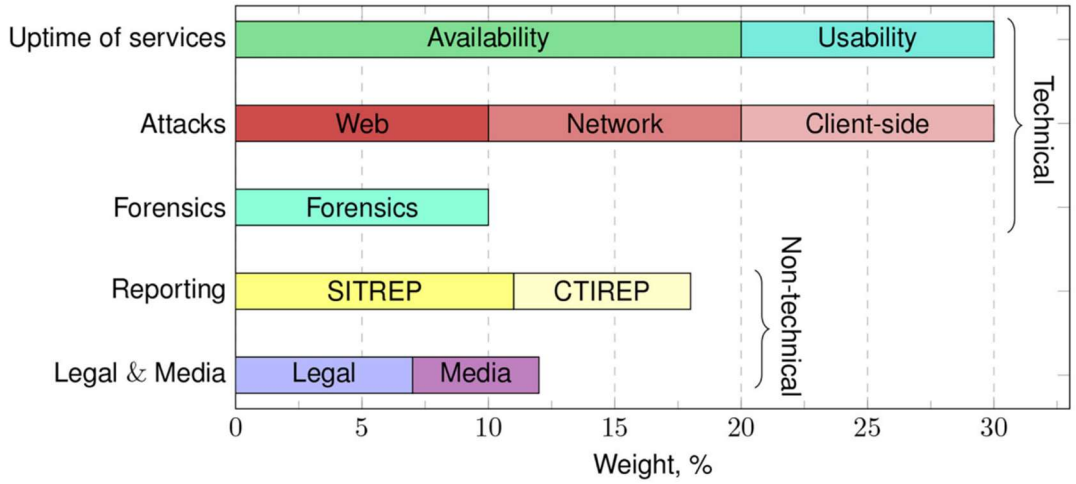
简单的评分可以评估已完成的任务与完成任务的时间，快速响应的防御者和快速渗透的攻击者可以获得额外的分数。大规模网络安全防御演习中的任务多种多样，线性评分机制很可能存在较大缺陷。尽管没有通用的解决方案可以让各参与方都满意，但 Locked Shields 的举办方根据多年的经验，总结了与评分机制有关的各种事务。有关评分机制的讨论，可以促进网络安全防御演习向更平衡、更实用、更有趣的方向发展。在评分机制的设计上，往往要考虑以下几点：

- 网络安全防御演习评分机制的趋势与挑战？
- 确保网络安全防御演习评分机制平衡性的因素有哪些？
- 如何在实践中确保这些影响平衡性的因素落实？

由于网络安全防御演习是十分复杂的，**简单的评分实际上无法完全反映参与者的实际能力，不同的目标需要不同的评分方法。**过去的评分机制多基于 CTF 比赛

中，而不是场景更为复杂的红蓝对抗。CTF 比赛的评分往往是基于机密性、完整性与可用性的，但网络安全防御演习的评估需要更多参数，例如成功缓解攻击的数量、信息共享质量等。

Locked Shields 的评分机制包含技术方面与非技术方面，权重（2022 年）大致如下所示：



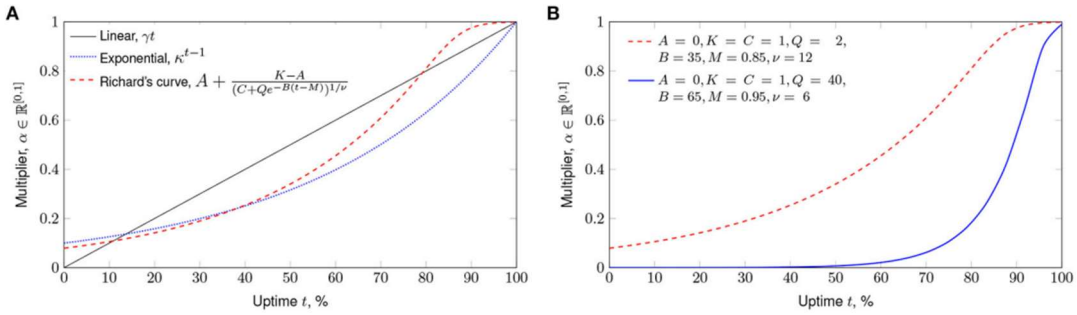
技术能力当然是最重要的，大约占到七成的分数，技术能力侧重于攻击、数字取证等。非技术能力侧重于“软技能”，例如口头与书面沟通能力等。但**不要认为非技术能力中毫无“技术”含量，例如蓝队提交的技术报告中会有法律分析的部分。**

蓝队有 1 天的时间熟悉相关网络环境，然后会面临 2 天的红队攻击。**Locked Shields 的参与团队都必须互相协作，而且必须与其他团队协作行动。**例如，负责电力基础设施的蓝队要保证配电设施的开放。

与服务正常运行相关的总分与攻击总分相等，服务无法正常运行通常是由于蓝队的过度保护造成的。最初的攻击通常针对机密性与完整性，将针对可用性的攻击留到最后。

服务正常运行类分数

与服务正常运行相关的分数共有两部分：**Availability 与 Usability。前者由机器自动进行检查，评分会考虑机器的重要性。后者由用户模拟团队中的模拟用户进行检查**，用户无法访问时会提交工单，工单未解决的时间越长，扣的分越多。2021 年开始，**Locked Shields 使用 Richard 曲线来衡量服务正常运行评分。**由于其与实际用户感知类似，最开始用户不太介意，随后耐心会被快速耗尽，最后用户不再关心服务是否恢复。



攻击类分数

与攻击相关的分数共有三部分：Web、网络与客户端。Web 攻击主要针对应用层漏洞，例如 Web 表单中存在未过滤的字段。网络攻击主要针对网络层进行攻击，例如针对防火墙中存在缺陷的 IPv6 规则。客户端攻击主要针对与人有关的威胁，例如用户模拟团队中的模拟用户会上传文件或者执行文件。

数字取证类分数

通过数字取证工具对事件进行调查，在提供的证据文件中查找恶意活动的痕迹。

报告类分数

报告用于向管理人员、决策人员通报网络状况，包括威胁、状态、关键事件、攻击者与技术挑战。**报告主要分为两部分：SITREP（情况报告）与 CTIREP（网络威胁情报报告）。**

法律与媒体类分数

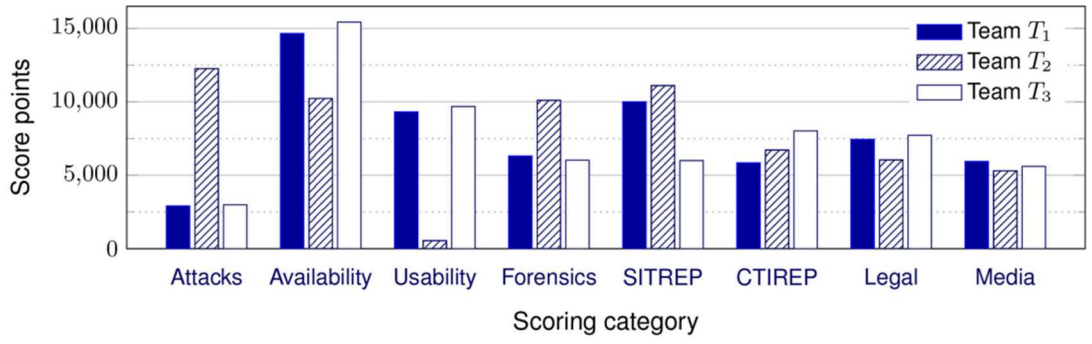
法律与媒体也十分重要，可以避免蓝队仓促采取可能扩大危机的措施。**蓝队必须要考虑立法与沟通上的问题，还需要在媒体上作出回应与解释。**

不同团队的任务不同，使用 12 分制李克特量表来兼容多种情况。

根据参与者的反馈，**与评分制度有关的反馈中 40% 都关注清晰度与透明度。**他们觉得并非所有所有攻击都经过充分论证，并且更清晰的反馈可以使参与者快速了解哪些人做对了、哪些人做错了。

Theme	Percentage
Clarity/Transparency	42
Feedback/Justification for scoring	26
System Set-up/Technical issues	16
Visibility/Comparability to others	6
Proportionality	3
“Game-ism”	3
Visualization	3

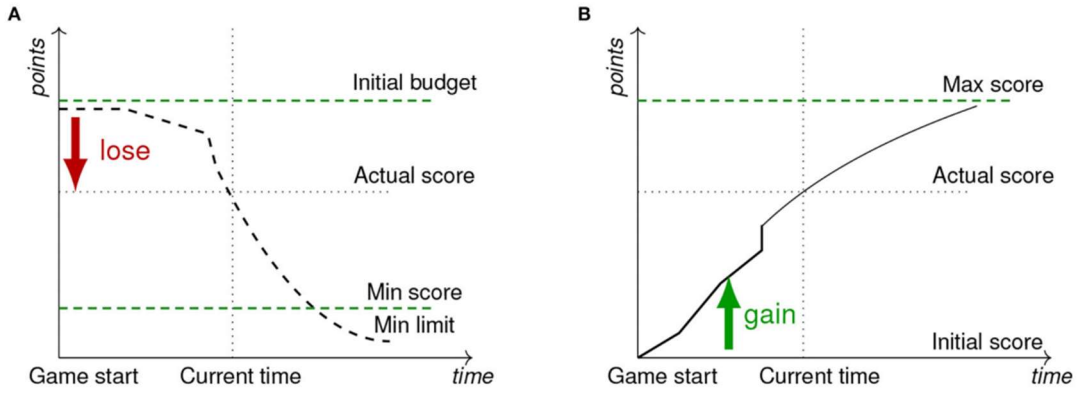
例如三支得分相近的蓝队具体如下所示，T1 与 T3 在确保服务正常运行上得分更高，但 T2 防御红队攻击则更出色。**可能就是由于其采取的防御措施，导致服务正常运行时间受到影响。**



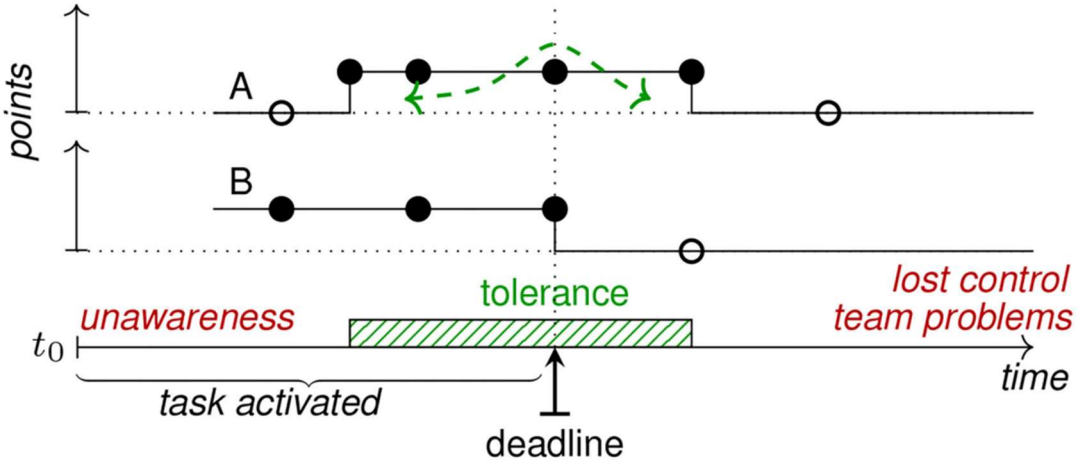
在满分 3200 分中获得 1600 分的表现，远不如满分 360 分中获得 350 分。**必须结合系统的优先级与重要性来评估分数。**在高价值任务中获取一半的分数，

但守住了低价值任务，这可能表明团队的优先级存在问题，或者严重缺乏某些能力。

评分的设计可以是递增或者递减的，尽管数学上一样，但**往往认为对损失的厌恶比潜在收益更能激励参与者。**



有一些任务会被设计成要在指定的截止日期前提交，评分中需要考虑时间带来的紧迫感。当然，由于是团队合作，必须允许存在一定区间的弹性来容忍延迟。此类任务的评估也有两种方法，如下所示：



评分总是很难的。例如如果两个队伍的危机沟通能力类似，但技术能力存在差别。一方在保护网络安全方面表现良好，另一方则明显较差。前者可能没有机会展示其危机沟通能力，这种内在的联系难以体现。

系统性、科学性的评分制度是为了衡量和评估差距与不足，为了考试而应试绝不可取。参与 Locked Shields 的人员都有共识：“**这应该是一次训练，不是一场计分比赛**”。倘若初心和理想越走越远，耗费了无数资源和人力的大考无法带来教训和养分，也许不可避免地要在一地鸡毛中越来越卷，这可能是所有人都不乐见的。