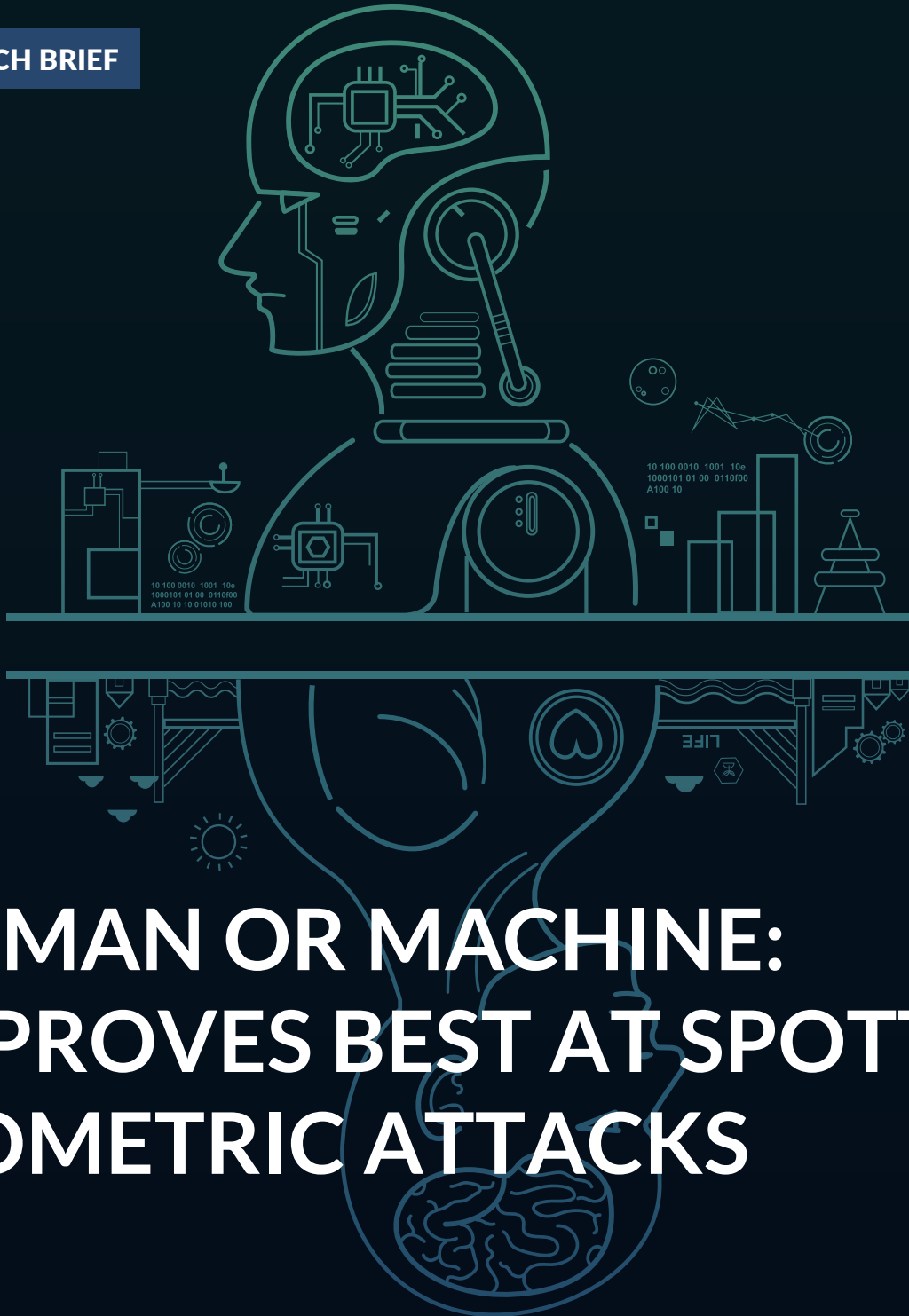


RESEARCH BRIEF



HUMAN OR MACHINE: AI PROVES BEST AT SPOTTING BIOMETRIC ATTACKS





INTRODUCTION

There are numerous news articles about the ability of machines, particularly AI-powered machines, to perform human jobs. In some cases, it makes sense. We all understand that a robot on an assembly line can work faster, more precisely, and perform repetitive tasks without breaks. But what about more complex situations where strategy and decision-making skills are required?

In 1996 an IBM supercomputer called Deep Blue played the world's reigning chess champion, Garry Kasparov. Garry won. A year later there was a rematch and Deep Blue became the first computer system to defeat a reigning champion in a chess match under standard tournament time controls. [Deep Blue](#) was soon surpassed by chess machines like Alpha Zero, which uses neural networks and deep learning to master the game. Exactly how good are these machines? Excellent, considering the last recorded instance of a human beating a computer at chess was in 2005.

Modern advances in Artificial Intelligence elevate the person vs. machine discussion to new levels. Often, humans don't believe that machines can outperform them. In specific scenarios, such as those requiring emotional intelligence, humans have the advantage. However, AI is valuable when performing particular decision-making tasks – many of which are resource-intensive for humans.

One such task is face recognition, and in this category falls a technology called **facial liveness detection**.

Facial liveness detection protects face biometric systems from spoofing attacks. It does this by determining if the person in front of the camera is present (live) or someone presenting a printed photo, video, digital image, or using a mask to trick the system into thinking it sees a real person.

With large organizations using biometric systems to process millions of face checks monthly, a leading advantage of AI liveness detection is the reduced accuracy burden on humans. Another advantage of AI is faster response times for the customer. But these advantages aside, which is better at the job, the human or AI?

Despite major advances in the accuracy and usability of liveness detection, many companies still use human analysts to manually verify the liveness of an applicant when onboarding new customers for an account. Others use a hybrid approach, with a significant portion of incoming images still going to humans for review.

Discerning between a selfie and a live selfie may seem to be a simple task. One might even assume that people are better at liveness detection than computers. In its experiences, the team at ID R&D found this was not the case. To further prove the value of an automated AI approach, ID R&D conducted a formal experiment to help determine which is better at liveness detection, people or machines?






The findings of the experiment follow.



MACHINE ACCURACY IN FACIAL LIVENESS DETECTION

The study used ID R&D's IDLive® Face product as the automated liveness detection system. This algorithm can distinguish between the presence of a live person in front of a camera and a spoof attack. IDLive Face uses a single image to evaluate liveness.

Fraudsters use the following types of attacks.

PRINTED PHOTO	The camera is presented with a printed photograph of the face.	
DISPLAY ATTACK	The camera is presented with a monitor display of a face image.	
PRINTED CUTOUT	The camera is presented with a mask of another person's face cut out of paper or cardboard.	
2D MASK	The camera is presented with a person holding a photographic mask of another person's face cut out of paper or cardboard. The mask partially covers the face of the person.	
3D MASK	The camera is presented with a person wearing a 3D human-like face mask, including realistic silicone masks.	



The evaluation dataset included the entire range of attacks - printed photos, display attacks, etc. The total number of attack images in the dataset was 175,454. Table 1 contains a detailed breakdown of the images by type. The dataset also included bona fide, or “live”, images.

Table 1. Description of test data utilized for IDLive Face evaluation

TEST		AMOUNT
Attacks	Printed photo	7551 images
	Display attack	33722 images
	Printed Cutout	33757 images
	2D mask	33484 images
	3D mask	33497 images
Bona Fide		33443 images

Machine accuracy was assessed using industry-standard metrics for measuring the two types of errors that occur when determining liveness. The Attack Presentation Classification Error Rate (“APCER” describes the rate of errors in which a spoofed image is classified as real. This means a fraudster is able to get through. The Bona Fide Presentation Classification Error Rate (“BPCER” is the rate of errors in which a live image is classified as a spoof. This error rate does not expose the biometric system to fraud, but it does create inconvenience for valid people. These metrics have an inverse correlation and both are important.

Liveness systems operate with a trade-off between these two types of errors. In practice, things go wrong – particularly with solutions that require users to take an action to prove liveness, such as blink, smile, turn a head, or move a device. BPCER for these solutions may be as high as 35%, meaning users abandon the process. Most businesses do not want to reject that many valid users during a liveness check.

The algorithm used in the experiment is a passive liveness check with no user action required. BPCER is still a risk because some valid images may appear to be spoofs. So, passive liveness also has an operating point with a trade-off between BPCER and APCER. APCER can be zero, but BPCER may be more than 1%. Most companies prefer to calibrate for a BPCER that is as low as possible while keeping the APCER as close to zero as possible.



It is important to recognize that APCER and BPCER metrics depend on the data set. The data set used in this experiment is based on a broad selection of data captured across age, race, and gender. The spoofing attacks were developed by ID R&D experts who are skilled at creating challenging spoof images.

The algorithm used in the experiment is optimized for balance between APCER (the number of spoofs that get through) and BPCER (the number of live images that are incorrectly classified as spoofs). This balance is configurable based on the use case. Allowing higher BPCER enables zero APCER.

As shown in Table 2, the passive liveness detection machine algorithm achieved a 0% APCER across all attack types while maintaining a low BPCER of less than 1%. Will humans be able to beat this?

Table 2. APCER of IDLive Face broken down by attack type

ATTACK TYPE	2D MASK	3D MASK	PRINTED CUTOUT	PRINTED PHOTO	DISPLAY
APCER	0.00%	0.00%	0.00%	0.00%	0.00%

Table 3. BPCER of IDLive Face

CLASS	BONA FIDE
BPCER	0.95%



HUMAN ACCURACY IN FACIAL LIVENESS DETECTION

ID R&D used a crowdsourcing platform to estimate how accurate people are at determining facial liveness. Consumer participants were asked to determine the presence of a live person in front of the camera or an attack. To directly compare human performance with machine performance, the same test dataset was utilized. However, the dataset was down sampled to decrease the size for further manual processing. 10,455 images were randomly selected out of 175,454; therefore the proportion of classes was not changed.

Study participants received detailed instructions on how to complete the task. To ensure understanding, examples were included as part of a training step. People who successfully completed the training exercise were allowed to proceed to the primary study. People who demonstrated confusion during training were not asked to continue.

In total, 8,821 people participated in the study with a dataset comprising 10,455 test images. An average of 17 different people reviewed each image for a total of 177,735 completed tasks. Each task consisted of the participant viewing an image and manually labeling it as “live” or a “spoof.”

Table 4. ACPER of human labeling broken down by type of attacks

ATTACK TYPE	2D MASK	3D MASK	PRINTED CUTOUT	PRINTED PHOTO	DISPLAY
APCER	2.04%	2.35%	2.04%	30.34%	15.04%

Table 5. BCPER of human labeling

CLASS	BONA FIDE
BPCER	18.28%

Of the attacks manually labeled by humans, simple attack types like printed photos and display images have error rates far above zero, representing a significant threat to organizations relying on manual assessment of an image for liveness. Likewise, the rate of bona fide people getting rejected (BPCER) is high.

As shown in table 4, people performed worse at detecting liveness across all attack types, including printed photos, the most common and easiest type of biometric spoofing attack used by bad actors. Subsequently, the rate at which people incorrectly classified live images as spoofs was far higher than in the machine test.



AN ADDITIONAL STUDY ON HUMAN ACCURACY

In an additional experiment on human accuracy, ID R&D combined the responses of a group of people for each task, using the majority decision. This approach differs from the prior human study, which recorded the decisions of participants individually. The experiment was based on the assumption that the accuracy of two people is statistically higher than the accuracy of one person, and the accuracy of a large group of people is higher than the accuracy of two. The size of the group used in the experiment was 17 people. Therefore each task response was based on a majority decision of 17 humans. As the number of people increases, the accuracy of the group response approximates the limit of human ability to differentiate bona fide images vs attacks.

Shown below in Tables 6 and 7, using the majority decision versus individual decisions resulted in lower error rates across all categories. This corroborates the assumption that the decision of multiple people is more accurate than any single person. However, even 17 humans could not beat the machine.

Table 6. ACPER of manual labeling (majority decision of 17 Users) broken down on attack type

ATTACK TYPE	2D MASK	3D MASK	PRINTED CUTOUT	PRINTED PHOTO	DISPLAY
APCER	0.15%	0.05%	0.00%	27.47%	8.7%

Table 7. BCPER of manual labeling (majority decision of 17 Users)

CLASS	BONA FIDE
BPCER	3.5%



THE RESULT: MACHINES WIN AT LIVENESS DETECTION

A summary of all findings is listed below in Table 8. As expected, the human eye easily detects 2D and 3D masks as well as printed cutouts. These types of attacks are obvious for most people, with only a small percentage of attacks missed by humans and close to 0% error rate when using the majority responses across a group.

Printed photos were the most challenging attack for the human eye to categorize accurately. Machine vision is significantly more accurate than people at distinguishing between printed and live photos. Screen Replay attacks on a display also prove to be challenging for human detection. Even when using the group’s majority decision, APCER was still over 8% compared to 0% for machines.

Table 8

ERROR	ATTACK TYPE	HUMAN (Single User)	HUMAN ABILITY LIMIT ESTIMATE (Group Decision)	MACHINE (IDLIVE Face)
APCER	2D Mask	2.04%	0.15%	0.00%
	3D Mask	2.35%	0.05%	0.00%
	Printed Cutout	2.04%	0.00%	0.00%
	Printed Photo	30.34%	27.47%	0.00%
	Display	15.04%	8.7%	0.00%
BPCER		18.28%	3.5%	0.95%

Taking into consideration that printed photos and replays are the easiest and most common types of attacks, it is important to classify them accurately. In these scenarios machine vision clearly outperforms the human eye.

Table 8 shows accuracy; however, accuracy is not the only important criterion for a liveness solution. Speed is another. On average, it took the humans in our test 4.8 seconds per image to determine liveness. The IDLive Face average speed is below 0.5 seconds per image for a single CPU core. It can be even faster for machines if more than a single core is utilized for image processing.

The findings show that the machine is both more accurate and faster than one or more humans.



HOW TO USE THIS STUDY

ID R&D's intent with this study is for businesses to better understand how the right liveness detection solution stacks up against a team of human reviewers. With a clear picture of the accuracy of automated liveness detection, businesses can confidently shift their human resources to other tasks.

ACCURACY: The study shows that a facial liveness detection system such as IDLive Face is more accurate than individuals or groups of people at identifying biometric spoofs - and is significantly better at categorizing the most frequent types of attacks. Equally important, the automated system has a far better bona fide rejection rate. Bona fide rejects are arguably more expensive for the business as they require human review and often lead to friction and customer abandonment. What's not tested here is a highly trained employee. Perhaps a business may close the gap between human and machine with a team of highly trained employees; however, the benefits of using automation extend beyond accuracy alone.

SCALE: As remote account onboarding and authentication rapidly increases, so does the need for liveness detection. Using an automated solution significantly improves the ability to scale. An organization can scale by adding servers and won't need to worry about employee turnover and time to train new employees. The advantage of automation over humans in scaling up a liveness system is enormous.

SPEED: The automated liveness check is measured in milliseconds. Manual liveness checks during onboarding can take days if there is a backlog. In authentication use cases, every second counts. A manual check isn't feasible as users expect near-instantaneous access. A single frame passive liveness solution, on the other hand, can reach a liveness decision in less than a second.

COST: Automation reduces the cost of liveness detection and frees valuable human capital to focus on other roles.

SUMMARY

In the experiment, the average APCER for a human was 10.36% compared to 0% for the machine algorithm. Machines also outperformed humans on the important factor of BPCER with an error rate of less than 1% compared to more than 18%. When factoring in speed, cost and scalability of the user experience, AI provides a clear advantage in liveness detection.

Table 9

	HUMAN	MACHINE
Avg APCER Spoofs let in	10.36%	0%
Avg BPCER Live users kept out	18.28%	.95%
Time to Detect	4.8 seconds	0.5 seconds

If you would like to learn more, visit www.idrnd.ai