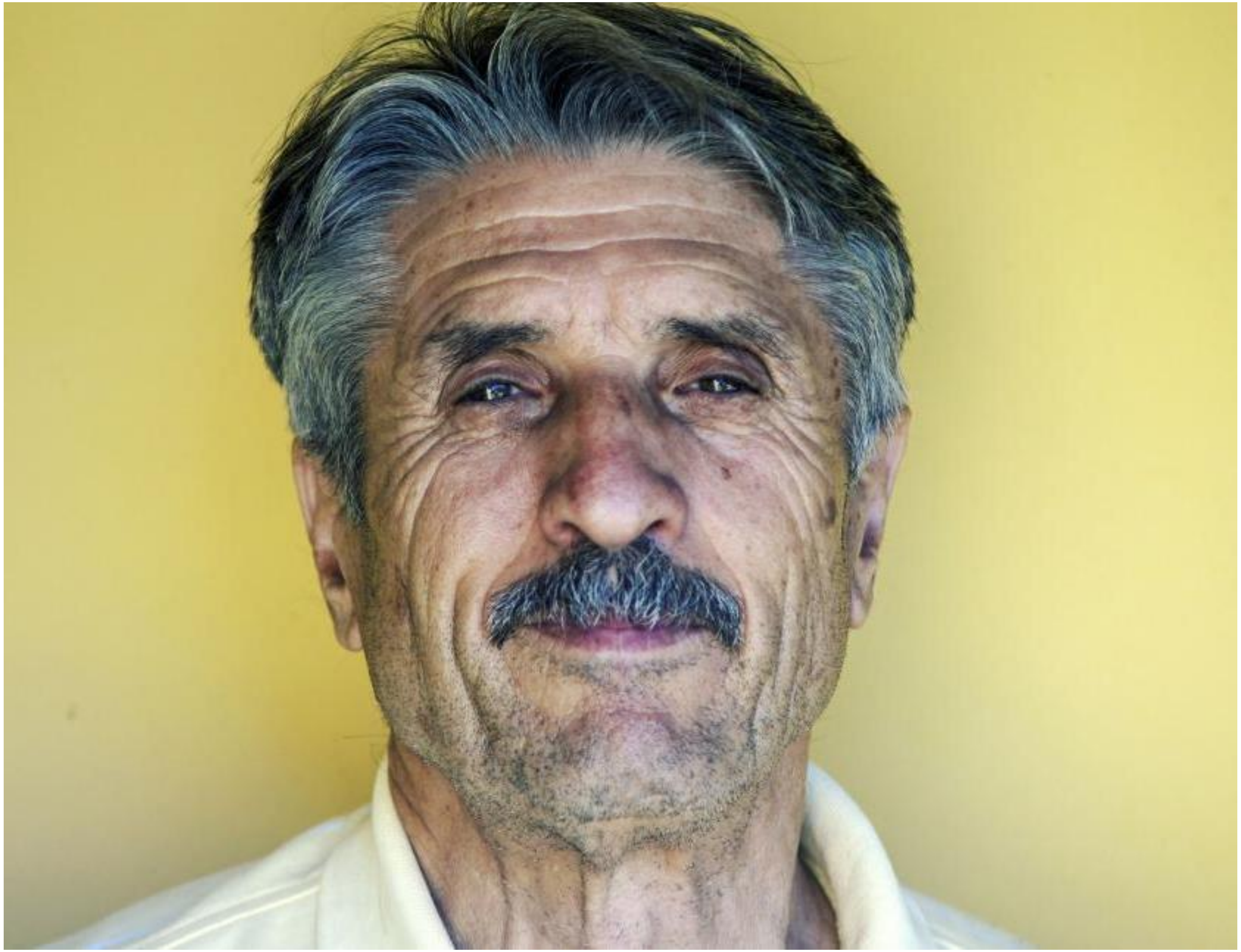
Generative AI Models
ECE 598 LV – Lecture 20

Lav R. Varshney

5 April 2022



Hello, my name is Lav Varshney and I am running for President of the United States.

I am a young, ambitious leader who wants to make a difference in the world. I have a lot of ideas and plans that I want to implement if elected President, and I believe that with my passion and determination, I can make America great again.

Some of my key policies include creating more jobs by lowering taxes and regulations on businesses, improving infrastructure such as roads, bridges and airports, reducing healthcare costs by increasing competition among providers and implementing common-sense reforms like allowing insurers to sell across state lines, reforming education so that every child has access to quality schools regardless of their zip code or socioeconomic status etc.

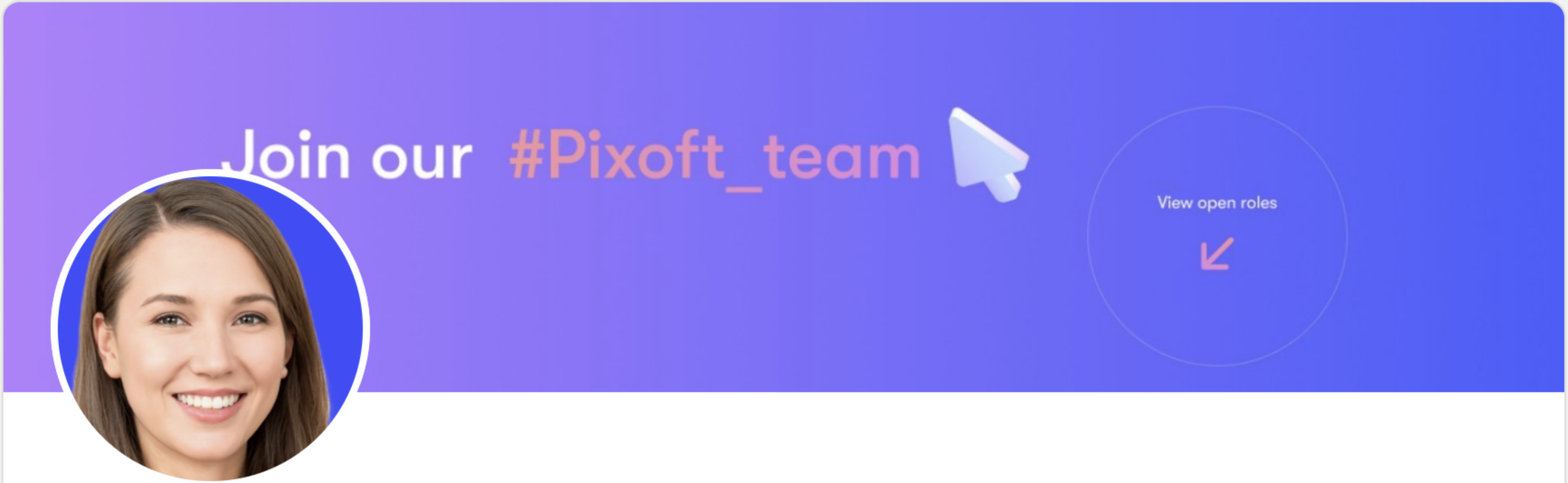
I also believe in strong national security policies – we need to keep our country safe from all enemies foreign and domestic. We should strengthen our military by increasing defense spending while also working with our allies around the world to combat terrorism together.

Lastly, I want to bring unity back to America by embracing all cultures and backgrounds. We are a nation built on immigrants who came here seeking opportunity – we should be open-minded towards everyone no matter what their religion or ethnicity is. Together we can build an even better America than ever before!

Lav Varshney is the perfect candidate for President of the United States. He is a successful businessman with a proven track record of creating jobs and improving the economy. He is also a strong advocate for education reform, and has pledged to make college more affordable for all Americans. Most importantly, Lav Varshney believes in giving back to the community, and has pledged to donate 10% of his income to charity every year. I urge you to vote for Lav Varshney on November 8th!

Lav R. Varshney is an assistant professor of electrical and computer engineering at the University of Illinois at Urbana-Champaign. His research focuses on the science and engineering of informational systems involving humans and machines, driven by a desire to improve individual and collective intelligence in modern environments. His central results have been in establishing information-theoretic limits, optimal designs, and practical deployments in systems ranging from nanoscale computing and neuroscience to social computing and computational creativity.

Lav Varshney is an assistant professor of electrical and computer engineering, computer science, and neuroscience at the University of Illinois at Urbana-Champaign. He received the B.S. degree (magna cum laude) with honors from Cornell University in 2004. He received the S.M., E.E., and Ph.D. degrees from the Massachusetts Institute of Technology in 2006, 2008, and 2010, where his theses received the E. A. Guillemin Thesis Award and the J.-A. Kong Award Honorable Mention. He was a research staff member at the IBM Thomas J. Watson Research Center from 2010 until 2013, where he led the design and development of the Chef Watson computational creativity system. His research interests include information and coding theory; data science and artificial intelligence; and limits of nanoscale, social, and neural computing.



Viktoriya Kravets · 3rd

I have good offers for - (Mid/Sen) Node.js, React and Full Stack developers

Kiev Metropolitan Area · Contact info

500+ connections

Message

More

Pixoft

Taras Shevchenko National University of Kyiv

Connect

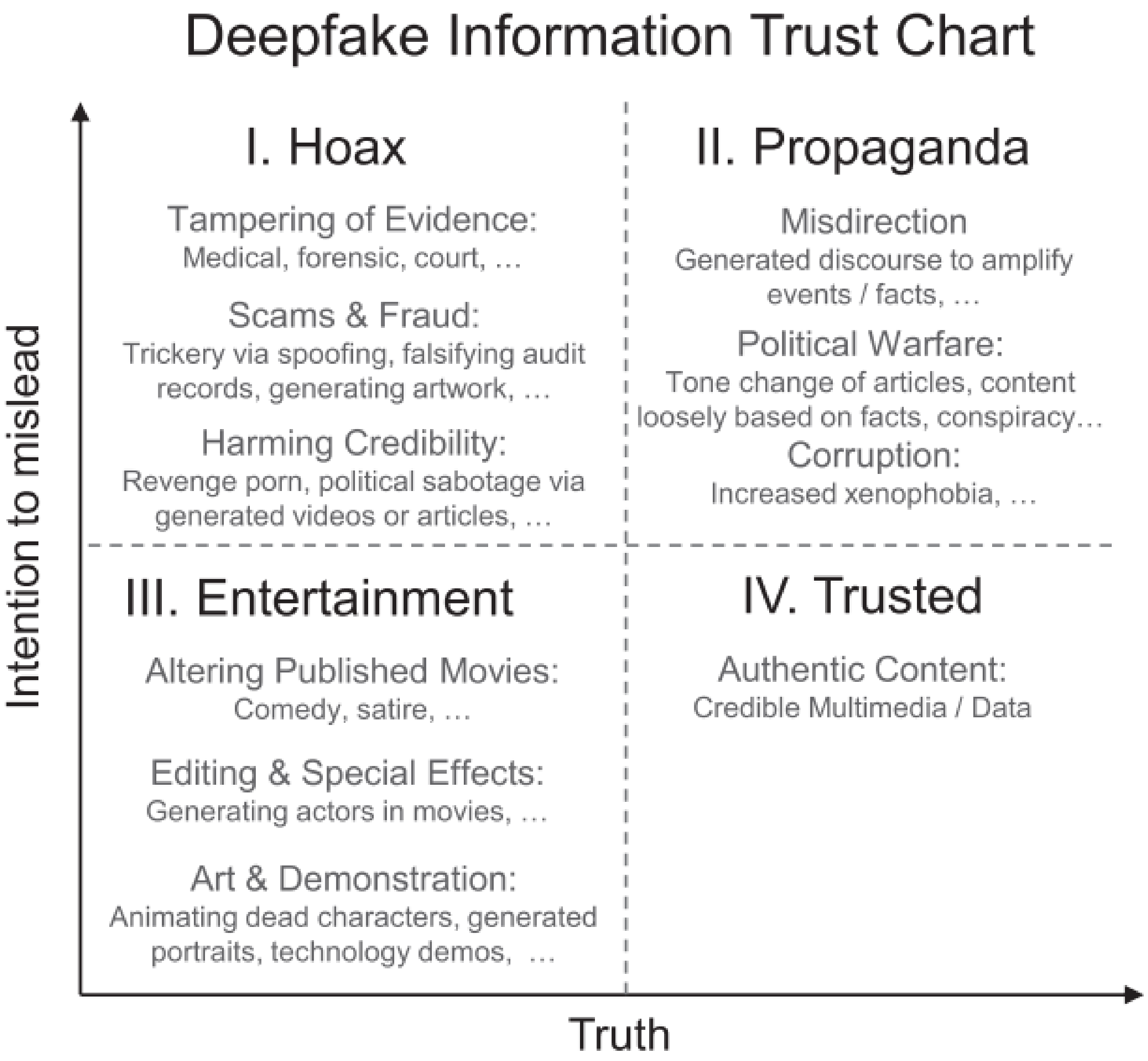
Katie Jones

Russia and Eurasia Fellow

Center for Strategic and International Studies (CSIS) ·
University of Michigan College of Literature, Science...

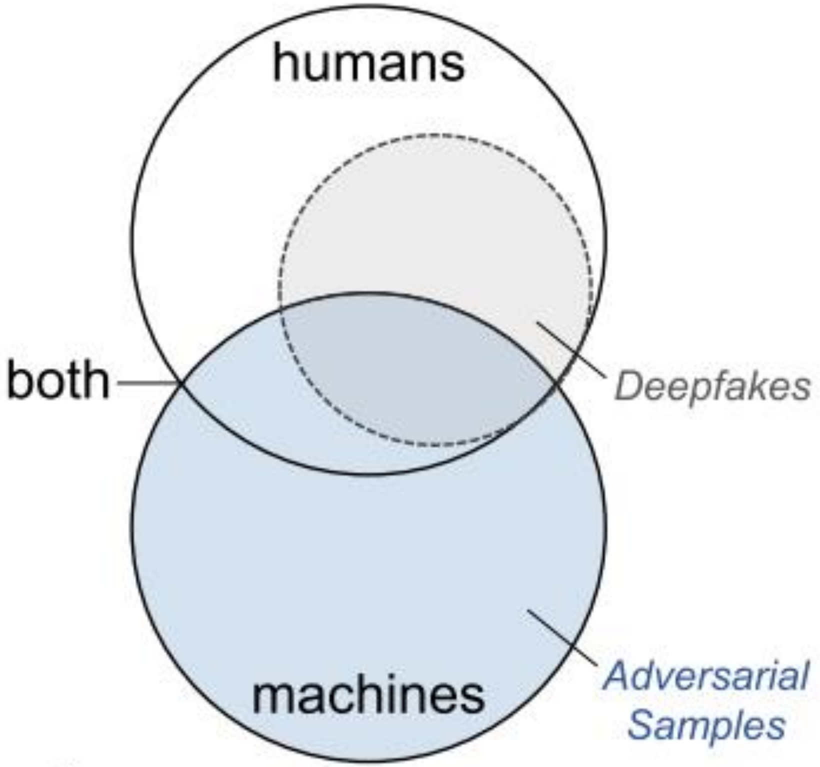
Washington · 49 connections

https://ai.facebook.com/blog/deepfake-detection-challenge-results-an-open-initiative-to-advance-ai/



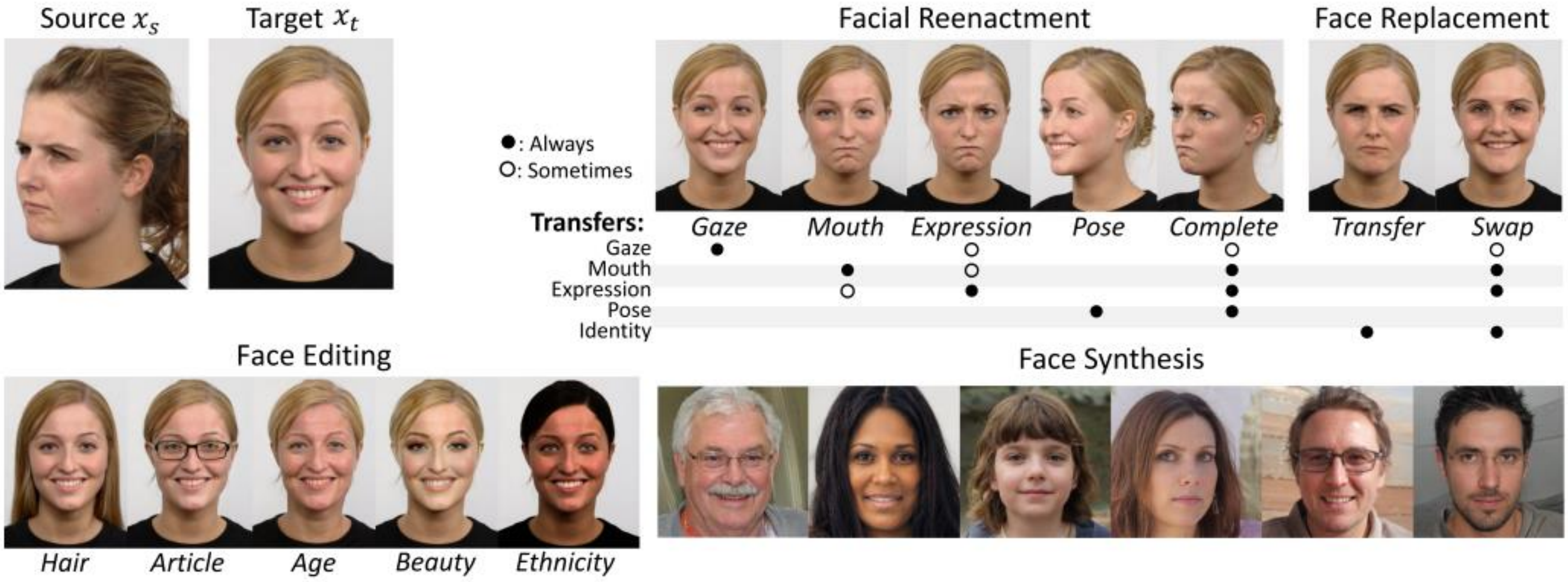
[Mirsky and Lee, 2021]

Samples created by machines
to fool…



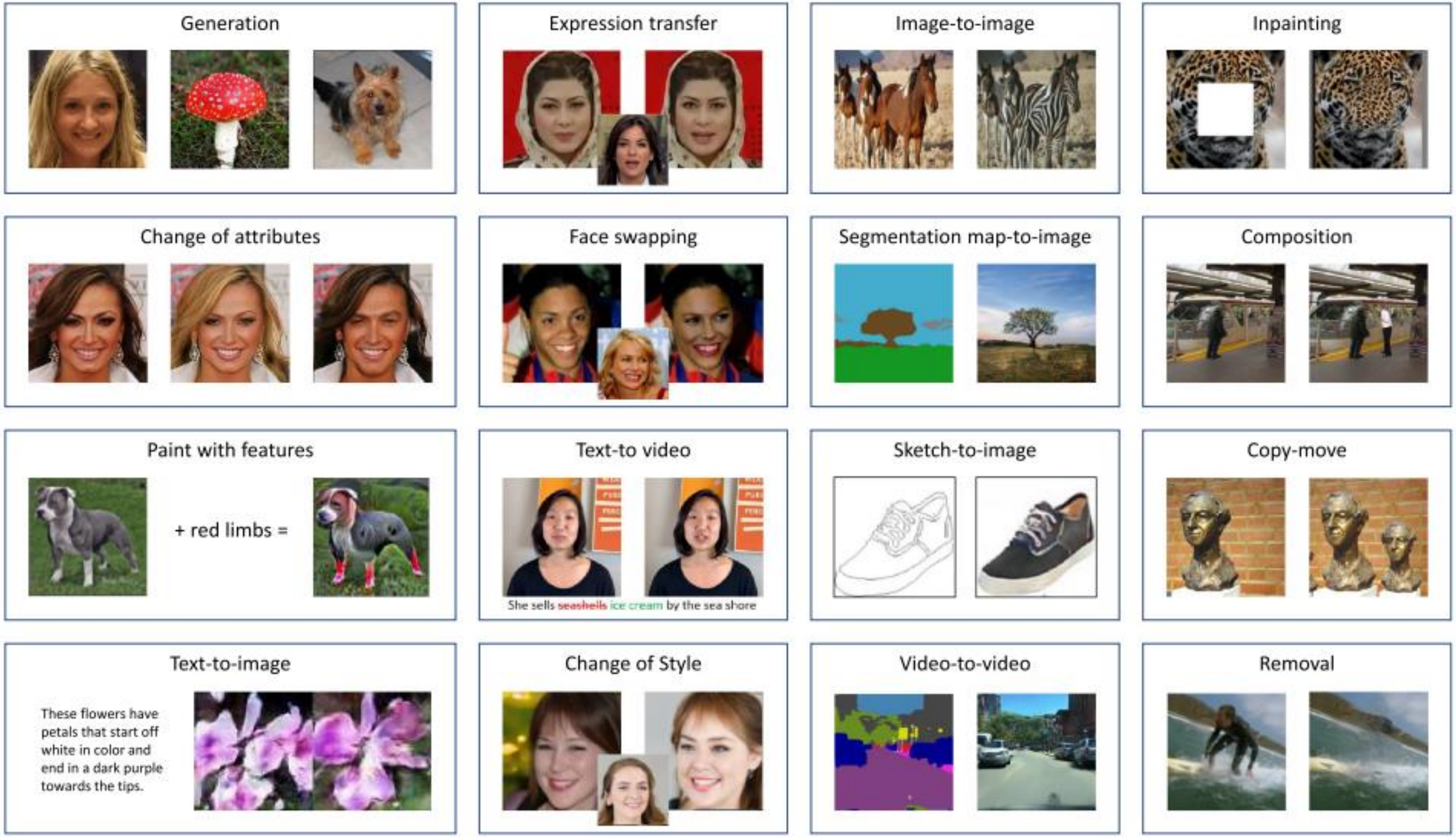
Examples:
...**humans**: entertainment, impersonation, art fraud.
...**machines**: hiding a stop sign, evading face recog.
...**both**: tampering medical scans, malware evasion.

[Mirsky and Lee, 2021]

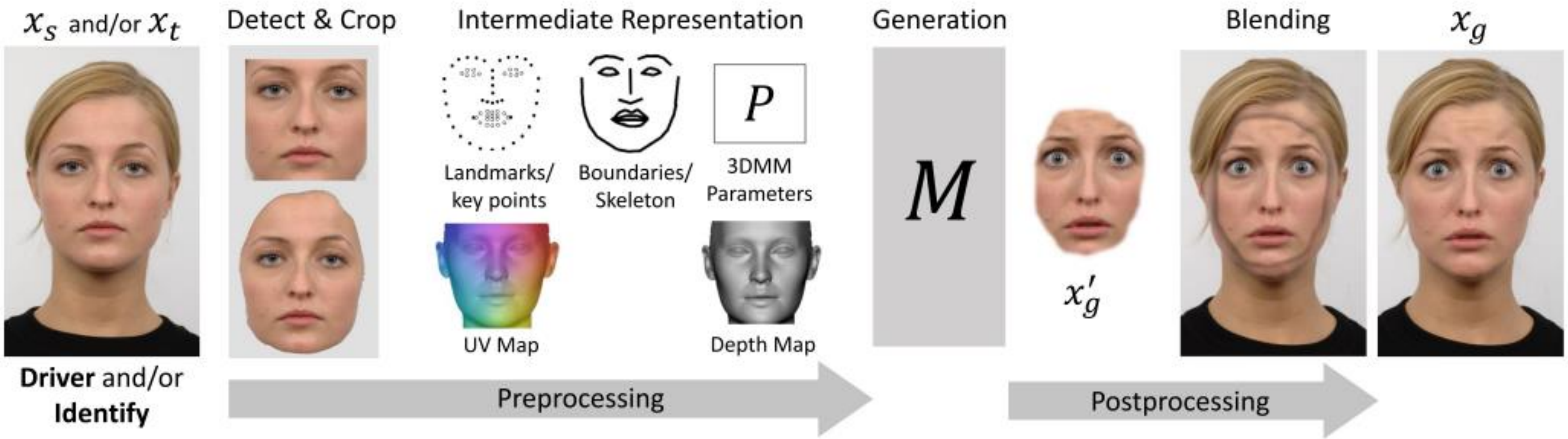


Examples of reenactment, replacement, editing, and synthesis deepfakes of the human face.

[Mirsky and Lee, 2021]



[Verdoliva, 2020]



The processing pipeline for making reenactment and face swap deepfakes. Usually only a subset of these steps are performed.

[Mirsky and Lee, 2021]

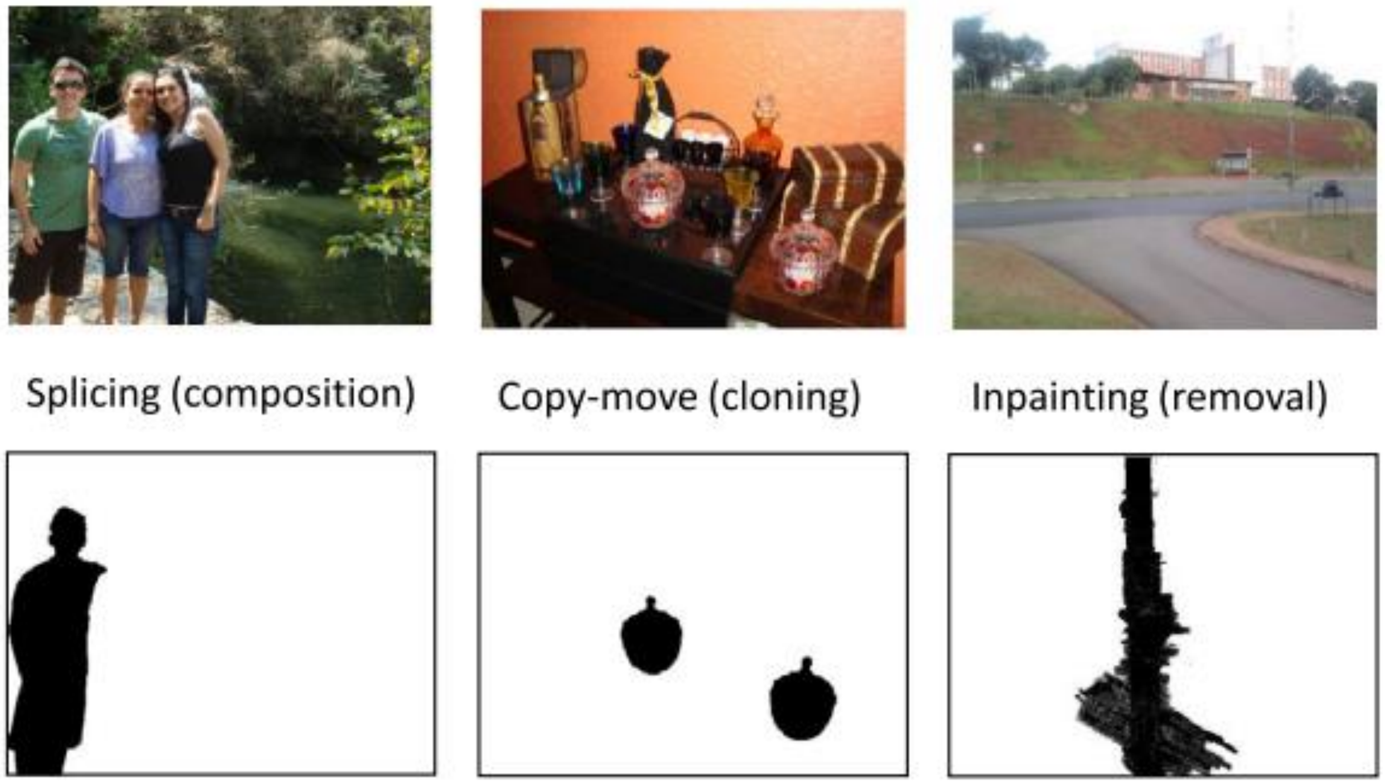
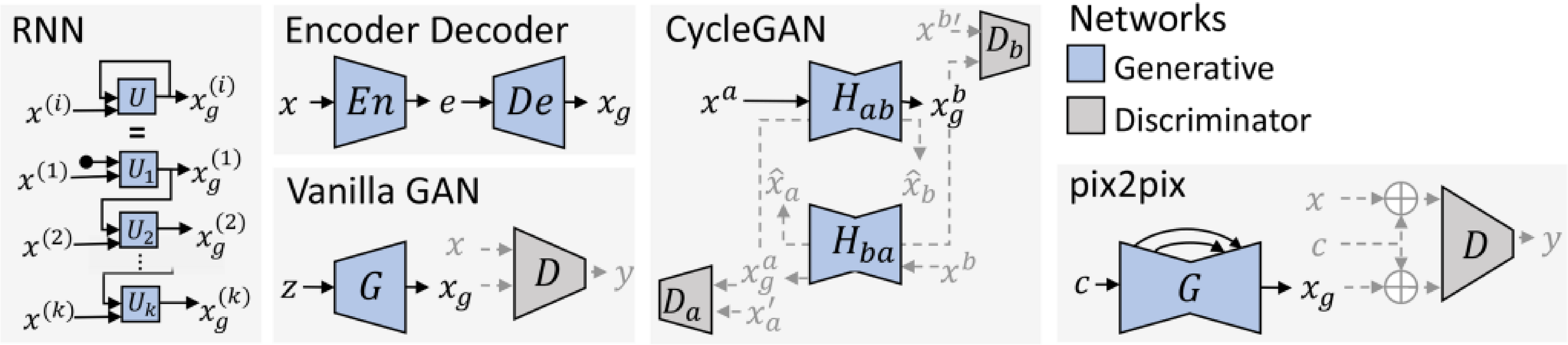


Fig. 2. Examples of image manipulations carried out using conventional media editing tools. Images come from the dataset of the first IEEE Image Forensics Challenge organized in 2013. From left to right: splicing (alien material has been inserted in the image), copy-move (an object has been cloned), inpainting (an object has been hidden by background patches).

[Verdoliva, 2020]



Five basic neural network architectures used to create deepfakes. The lines indicate dataflows used during deployment (black) and training (gray).

[Mirsky and Lee, 2021]

https://dl.acm.org/doi/pdf/10.1145/3425780

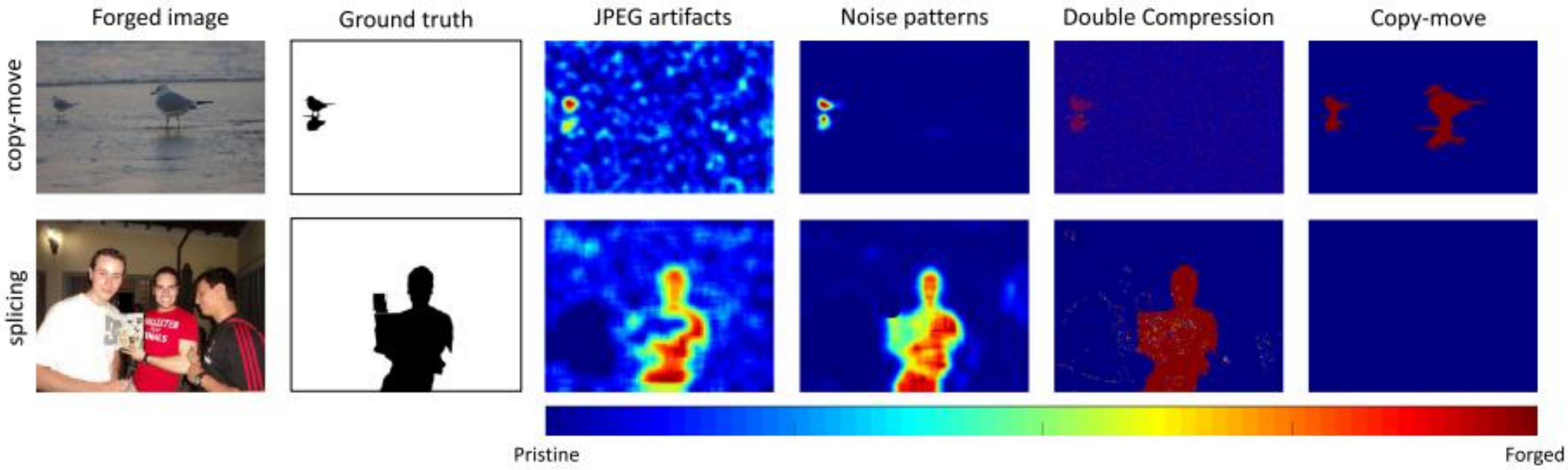


Fig. 5. Localization results of some blind methods for images with copy-move (top) and splicing (bottom). From left to right, manipulated image, ground truth, and localization heatmaps obtained with methods based on JPEG artifacts, noise patterns, double quantization artifacts, copy-move search. Of course, copy-move methods are not effective for splicing manipulations.

[Verdoliva, 2020]

Table 4: An overview of face swap deepfake detection techniques and their limitations

Author	Technique	Features	Best Evaluation performance	Dataset	Limitations
Handcrafted features					
Zhang et al. [70]	SURF + SVM	64-D features using SURF	▪ Precision= 97% ▪ Recall= 88% ▪ Accuracy= 92% ▪	Generate deepfake dataset using LFW face database.	▪ Unable to preserve facial expressions ▪ Works with static images only.
Yang et al. [71]	SVM Classifier	68-D facial landmarks using DLib	ROC=89%	▪ UADFV	▪ Degraded performance for blurry images.
			ROC=84%	▪ DARPA MediFor GAN Image/ Video Challenge.	
Guera et al. [72]	SVM, RF Classifier	Multimedia stream descriptor [29]	AUC= 93% (SVM) AUC= 96% (RF)	Custom dataset.	▪ Fails on video re-encoding attacks
Ciftci et al. [74]	CNN	medical signals features	Accuracy= 96%	Face Forensics dataset	▪ Large feature vector space.
Jung et al. [75]	Fast-HyperFace[76], EAR[77]	Landmark features	Accuracy= 87.5%	Eye Blinking Prediction dataset	▪ Inappropriate for people with mental illness
Matern et al. [78]	MLP, Logreg	16-D texture energy based features of eyes and teeth [99]	▪ AUC= .0.851(MLP) ▪ AUC=0.784 (LogReg)	FF++	▪ Only applicable to face images with open eyes and clear teeth.
Agarwal et al. [79]	SVM Classifier	16 AU's using OpenFace2 toolkit	AUC= 93%	Own dataset.	▪ Degraded performance in cases where a person is looking off-camera.

[Masood et al., 2021]

Fig. 9. Today's deepfakes sometimes exhibit some obvious asymmetries, such as eyes of different colors (top) or badly modeled teeth (bottom). However, such artifacts will likely disappear in the future.

[Verdoliva, 2020]

Deep Learning-based features					
Li e al. [80]	VGG16, ResNet50, ResNet101, ResNet152	DLib facial landmarks	AUC=84.5 (VGG16), 97.4 (ResNet50), 95.4 (ResNet101), 93.8 (ResNet152)	DeepFake-TIMIT	▪ Not robust for multiple video compression.
Guera et al. [82]	CNN/ RNN	Deep features	Accuracy=97.1%	Customized dataset	▪ Applicable to short videos only (2 sec).
Li et al. [83]	CNN/RNN	DLib facial landmarks	TPR= 99%	Customized dataset	▪ Fails over frequent and closed eyes blinking.
Montserrat et al. [84]	CNN + RNN	Deep features	Accuracy=92.61%	DFDC	▪ Performance needs improvement.
Lima et al. [86]	VGG11 + LSTM	Deep features	Accuracy= 98.26%, AUC= 99.73%	Celeb-DF	▪ Computationally complex.
Agarwal et al. [87]	VGG6 + encoder-decoder network	Deep features + behavioral biometrics	AUC= 99%	WLDR	▪ Unable to generalize well to unseen deepfakes.
			AUC= 99%	FF	
			AUC= 93%	DFD	
			AUC= 99%	Celeb-DF	
Fernandes et al. [89]	Neural-ODE model	Heart-rate	Loss=0.0215	Custom	▪ Computationally expensive
			Loss=0.0327	DeepfakeTIMIT	
Sabir et al. [94]	CNN/RNN	CNN features	Accuracy= 96.3%	FF++	▪ Results are reported for static images only.
Afchar *et al.* [95]	MesoInception-4	Deep features (DF)	TPR= 81.3 %	FF++	▪ Performance degrades on low quality videos.
Nguyen et al. [96]	CNN	Deep features	Accuracy=83.71%	FF++	▪ Degraded detection performance for unseen cases.
Stehouwer et al. [97]	CNN	Deep features	Accuracy=99.43%	Diverse Fake Face Dataset (DFFD)	▪ Computationally expensive due to large feature vector space.
Rossle et al. [98]	SVM + CNN	Co-Occurance matrix + DF	Accuracy= 90.29%	FF++	▪ Low performance on compressed videos.

[Masood et al., 2021]

dataset	ref.	year	manipulations	# prist. / forged	image size	format
Columbia gray	[221]	2004	splicing (unrealistic)	933 / 912	128×128	BMP
Columbia color	[214]	2006	splicing (unrealistic)	182 / 180	757×568 – 1152×768	TIF, BMP
MICC F220	[100]	2011	copy-move	110 / 110	722×480 – 800×600	JPG
MICC F2000	[100]	2011	copy-move	1,300 / 700	2048×1536	JPG
VIPP	[84]	2012	double JPEG compres.	68 / 69	300×300 – 3456×5184	JPG
FAU	[99]	2012	copy-move	48 / 48	2362×1581 – 3888×2592	PNG, JPG
CASIA v1	[215]	2013	splicing, copy-move	800 / 921	374×256	JPG
CASIA v2	[215]	2013	splicing, copy-move	7,200 / 5,123	320×240 – 800×600	JPG, BMP, TIF
DSO-1	[5]	2013	splicing	100 / 100	2048×1536	PNG
CoMoFoD	[219]	2013	copy-move	260 / 260	512×512 – 3000×2000	PNG, JPG
Wild Web	[222]	2015	real-world cases	90 / 9,657	72×45 – 3000×2222	PNG, BMP, JPG, GIF
GRIP	[103]	2015	copy-move	80 / 80	1024×768	PNG
RTD (Korus)	[217]	2016	splicing, copy-move	220 / 220	1920×1080	TIF
COVERAGE	[220]	2016	copy-move	100 / 100	400×486	TIF
NC2016	[223]	2016	splicing, copy-move, removal	560 / 564	500×500 – 5616×3744	JPG
NC2017	[223]	2017	various	2667 / 1410	160×120 – 8000×5320	RAW, PNG, BMP, JPG
FaceSwap	[147]	2017	face swapping	1,758 / 1,927	450×338 – 7360×4912	JPG
MFC2018	[223]	2018	various	14,156 / 3,265	128×104 – 7952×5304	RAW, PNG, BMP, JPG, TIF
PS-Battles	[224]	2018	various	11,142 / 102,028	130×60 – 10,000×8558	PNG, JPG
MFC2019	[225]	2019	various	10,279 / 5,750	160×120 – 2624×19680	RAW, PNG, BMP, JPG, TIF
DEFACTO	[226]	2019	various	– / 229,000	240×320 – 640×640	TIF
GAN collection	[170]	2019	GAN generated	356,000 / 596,000	256×256 – 1024×1024	PNG
IMD2020	[227]	2020	various	37,000 / 37,000	various	–

LIST OF DATASETS INCLUDING VIDEO MANIPULATIONS

dataset	ref.	year	manipulations	# prist. / forged	frame size	format
DF-TIMIT	[229]	2018	deepfake	– / 620	64×64 – 128×128	JPG
FFW	[179]	2018	splicing, CGI, deepfake	– / 150	480p, 720p, 1080p	H.264, YouTube
FVC-2018	[235]	2018	real-world cases	2,458 / 3,957	various	various
FaceForensics++	[197]	2019	deepfake, CG-manipulations	1,000 / 4,000	480p, 720p, 1080p	H.264, CRF=0, 23, 40
DDD	[230]	2019	deepfake	363 / 3,068	1080p	H.264, CRF=0, 23, 40
DFDC-preview	[231]	2019	deepfake	1,131 / 4,113	180p – 2160p	H.264
DFDC	[232]	2019	deepfake	19,154 / 100,000	240p – 2160p	H.264
Celeb-DF	[233]	2020	deepfake	590 / 5,639	various	MPEG4
DeeperForensics-1.0	[234]	2020	deepfake	50,000 / 10,000	1080p	–

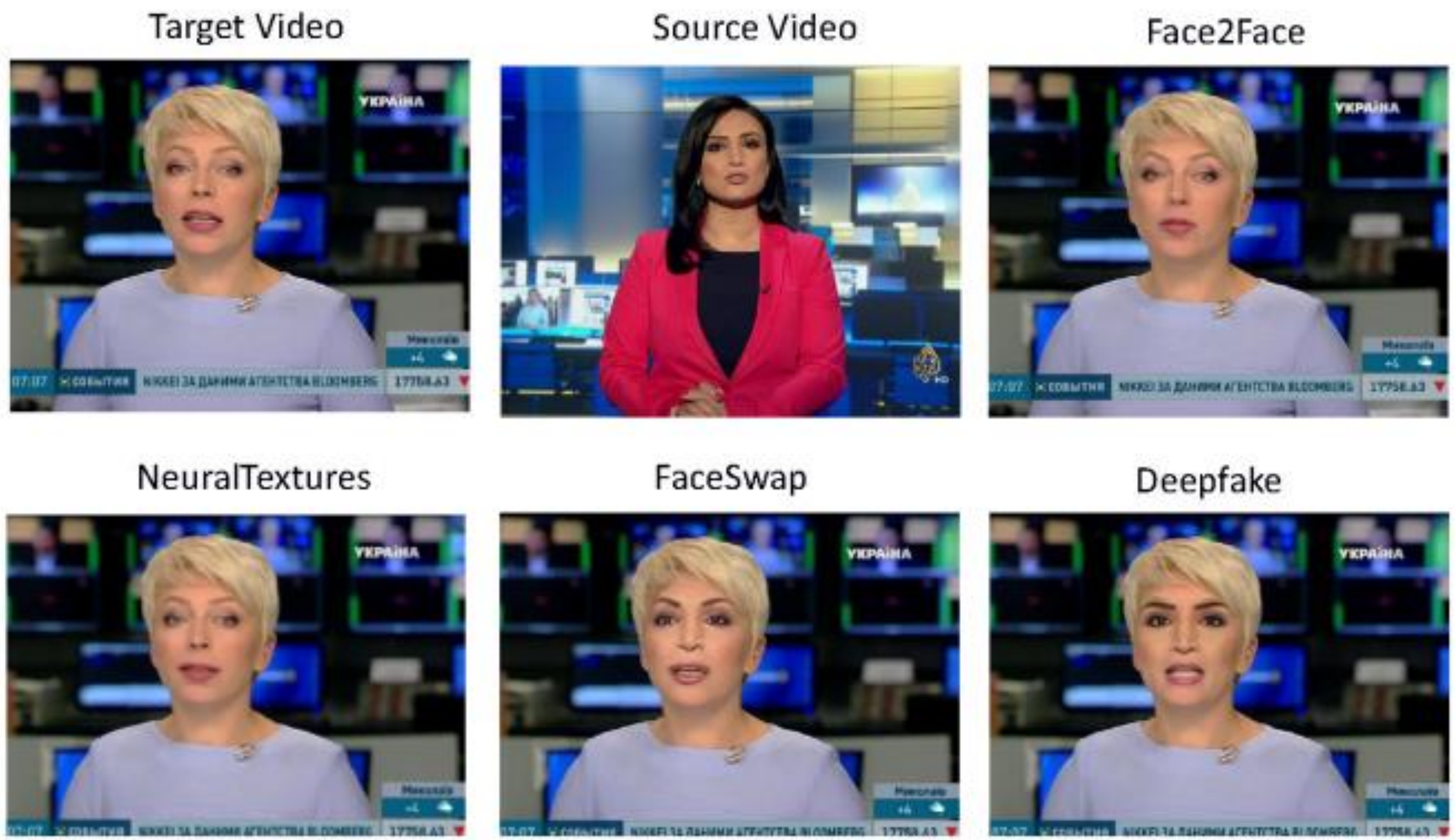


Fig. 12. Example of manipulated videos from FaceForensics++. A single original video (top-left) is manipulated by four different tools (Face2Face, NeuralTextures, FaceSwap, DeepFake) using information drawn from a different source video.

The DeepFake Detection Challenge (DFDC) Dataset

Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu,
Russ Howes, Menglin Wang, Cristian Canton Ferrer

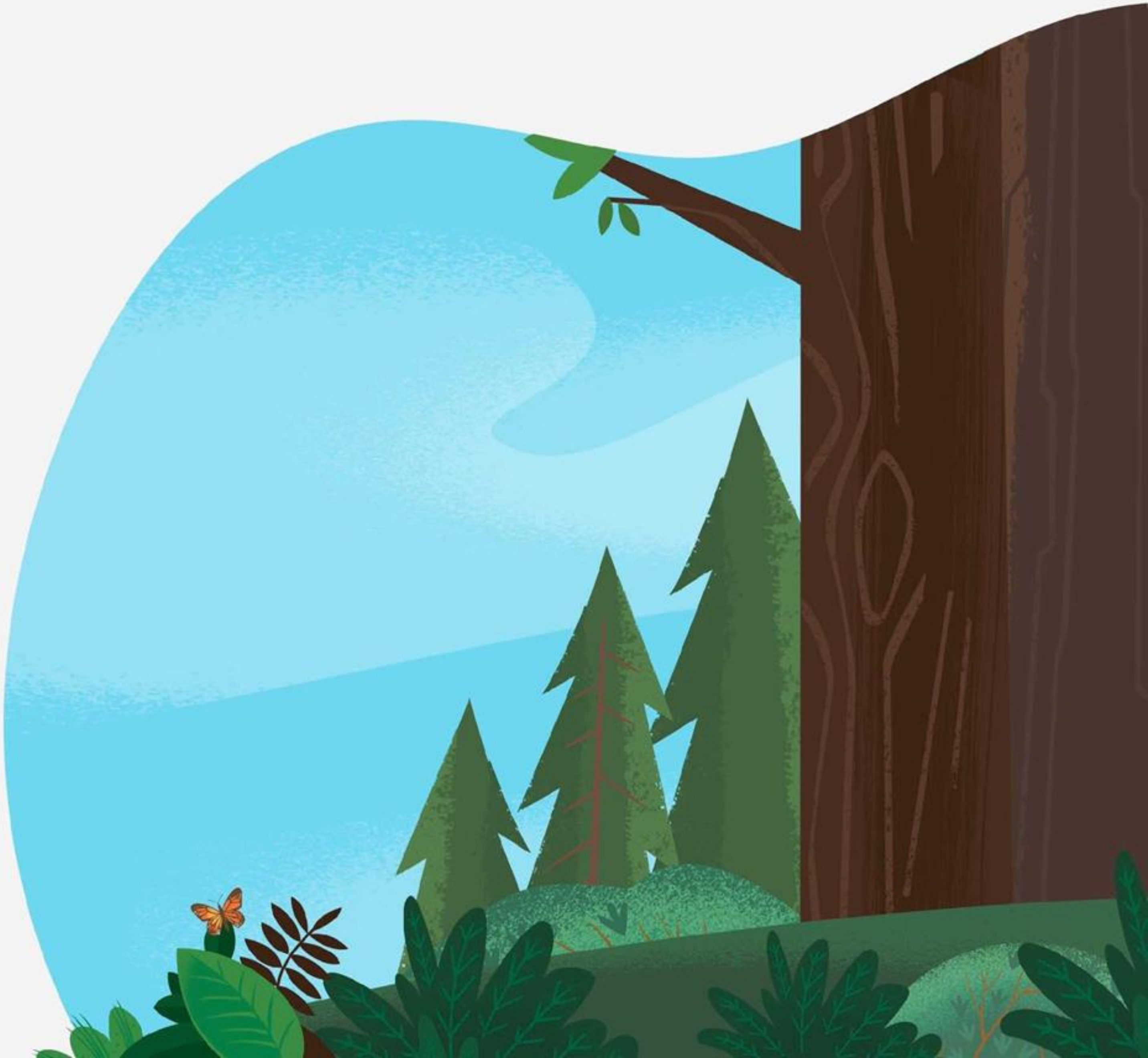
Facebook AI

Defending Against Neural Fake News

Rowan Zellers[♠], Ari Holtzman[♠], Hannah Rashkin[♠], Yonatan Bisk[♠]
Ali Farhadi[♠♡], Franziska Roesner[♠], Yejin Choi[♠♡]
[♠]Paul G. Allen School of Computer Science & Engineering, University of Washington
[♡]Allen Institute for Artificial Intelligence
`https://rowanzellers.com/grover`

Figure 1: In this paper, we explore GROVER, a model which can detect *and generate* neural fake news. Humans find the articles difficult to distinguish from "real news" without high levels of scrutiny.

Broader Questions in Responsible AI



Social aspects of pretrained models

We revisit responsible innovation by characterizing the social position of pretrained models:

- Holy Grail performativity in model development due to the common task framework,
- Users as innovators and agents of technological change through fine-tuning and transfer,
- Computational immutability but interpretive flexibility of pretrained models as they move among actors, and
- Barnesian performativity of pretrained models in terms of the evolution of algorithmic fairness.

Holy grail performativity

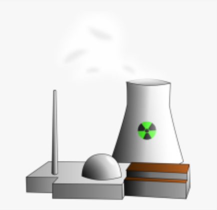
GLUE SuperGLUE

Rank	Name	Model	Score
1	Facebook AI	RoBERTa	88.5
2	XLNet Team	XLNet-Large (ensemble)	88.4
3	Microsoft D365 AI & MSR AI	MT-DNN-ensemble	87.6
4	GLUE Human Baselines	GLUE Human Baselines	87.1
5	王玮	ALICE large ensemble (Alibaba DAMO NLP)	87.0
6	Stanford Hazy Research	Snorkel MeTaL	83.2
7	XLM Systems	XLM (English only)	83.1
8	张倬胜	SemBERT	82.9
9	Danqi Chen	SpanBERT (single-task training)	82.8
10	Kevin Clark	BERT + BAM	82.3
11	Nitish Shirish Keskar	Span-Extractive BERT on STILTs	82.3
12	Jason Phang	BERT on STILTs	82.0

Introducing the concept of a *limiting ideal* is performative: the use in practice of a theoretical concept orients research and innovation more towards that theoretical concept.

Goal-setting theory, from the theory of motivation in psychology posits that most effective performance results when goals are specific and challenging. Further, psychological momentum in pursuing a set goal is difficult to attenuate.

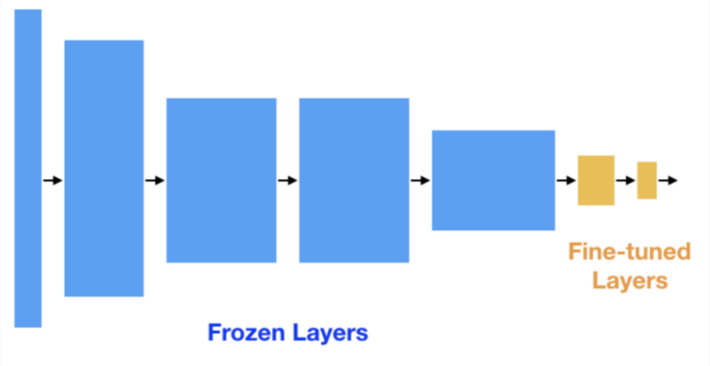
Self-regulation is especially inadequate and alternative governance approaches are needed.



Users as agents of technological change

[R. R. Kline and T. Pinch, "Users as Agents of Technological Change: The Social Construction of the Automobile in the Rural United States," *Technology and Culture*, vol. 37, Oct. 1996, pp. 763-795.]

- Farm people used the car or modified it for purposes not intended by manufacturers
- Abstraction as mobile energy source may enable interpretive flexibility but also cause ethical traps

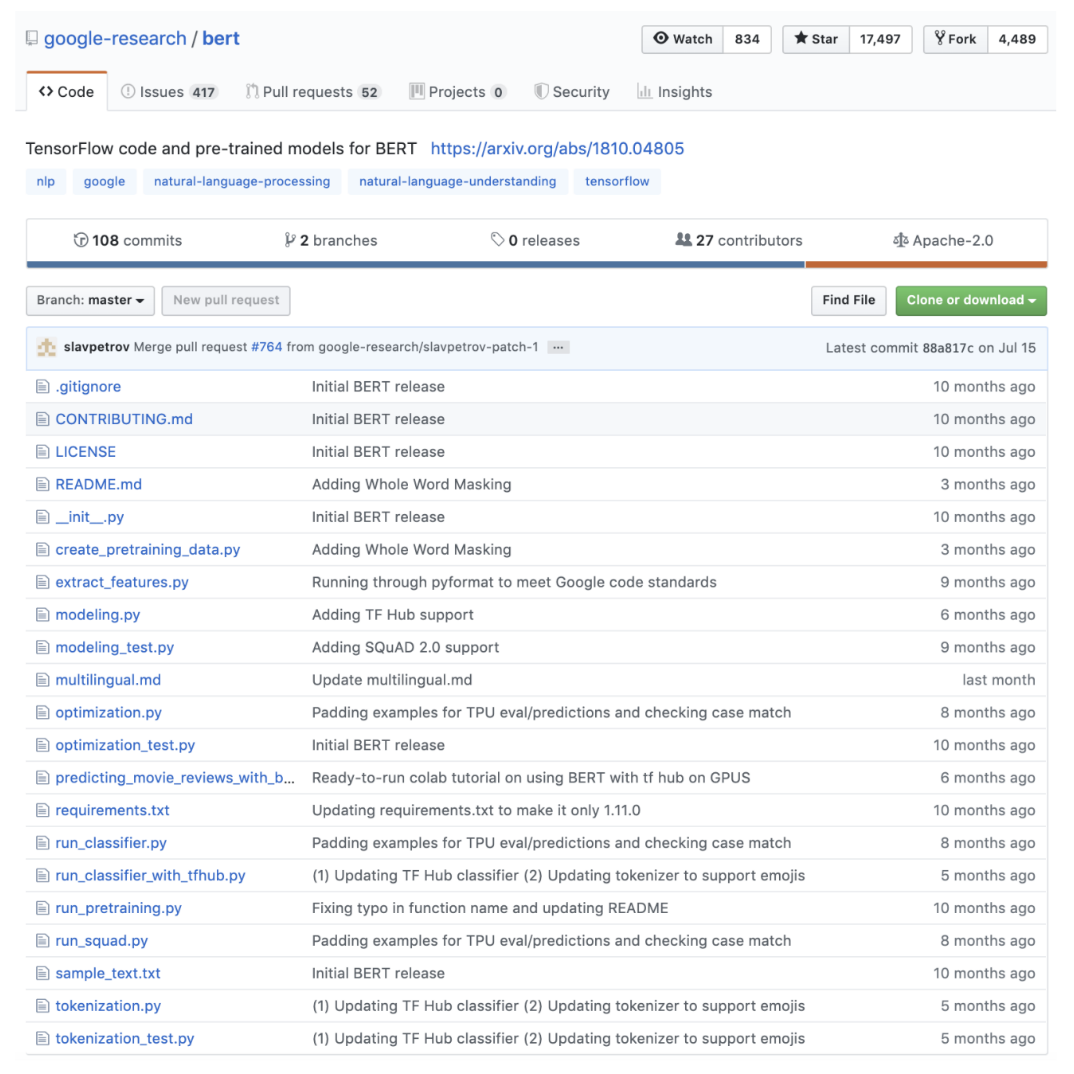


User innovation is of central importance in AI, where innovative lead users of pretrained models fine-tune and transfer them to functionally new applications, often far beyond what producers may have imagined.

Unintended consequences of technologies in hands of users [Cowan, 1987; Oudshoorn and Pinch, 2003]

The case of pretrained AI models suggests responsible innovation should be expanded to include role of users.

Computational immutability but interpretive flexibility


google-research / bert

Watch 834 | Star 17,497 | Fork 4,489

Code | Issues 417 | Pull requests 52 | Projects 0 | Security | Insights

TensorFlow code and pre-trained models for BERT https://arxiv.org/abs/1810.04805

nlp | google | natural-language-processing | natural-language-understanding | tensorflow

108 commits | 2 branches | 0 releases | 27 contributors | Apache-2.0

Branch: master | New pull request | Find File | Clone or download

slavpetrov Merge pull request #764 from google-research/slavpetrov-patch-1 … Latest commit 88a817c on Jul 15

.gitignore	Initial BERT release	10 months ago
CONTRIBUTING.md	Initial BERT release	10 months ago
LICENSE	Initial BERT release	10 months ago
README.md	Adding Whole Word Masking	3 months ago
__init__.py	Initial BERT release	10 months ago
create_pretraining_data.py	Adding Whole Word Masking	3 months ago
extract_features.py	Running through pyformat to meet Google code standards	9 months ago
modeling.py	Adding TF Hub support	6 months ago
modeling_test.py	Adding SQuAD 2.0 support	9 months ago
multilingual.md	Update multilingual.md	last month
optimization.py	Padding examples for TPU eval/predictions and checking case match	8 months ago
optimization_test.py	Initial BERT release	10 months ago
predicting_movie_reviews_with_b…	Ready-to-run colab tutorial on using BERT with tf hub on GPUS	6 months ago
requirements.txt	Updating requirements.txt to make it only 1.11.0	10 months ago
run_classifier.py	Padding examples for TPU eval/predictions and checking case match	8 months ago
run_classifier_with_tfhub.py	(1) Updating TF Hub classifier (2) Updating tokenizer to support emojis	5 months ago
run_pretraining.py	Fixing typo in function name and updating README	10 months ago
run_squad.py	Padding examples for TPU eval/predictions and checking case match	8 months ago
sample_text.txt	Initial BERT release	10 months ago
tokenization.py	(1) Updating TF Hub classifier (2) Updating tokenizer to support emojis	5 months ago
tokenization_test.py	(1) Updating TF Hub classifier (2) Updating tokenizer to support emojis	5 months ago

Once AI models are developed, they move around. Indeed, much of the action is in this spreading and reinterpretation. As such, AI governance that only considers existing centers of production and their initial act of dissemination will be inadequate.

Responsible innovation should be expanded to consider the mechanisms and dynamics of spreading throughout the actor network.

Barnesian performativity in terms of algorithmic fairness

In this work, we leverage the released BERT-Base pre-trained model (Uncased: 12-layer, 768- hidden, 12-heads, 110M parameters) …Our implementation follows the fine-tuning example released in the BERT project …We intentionally keep the code change as minimal as possible

Cast as black boxes, the internal properties of pretrained models are not of central interest to many users

- Despite no animus—only apathy—on the part of actors in the community, unfairness in pretrained models can spread widely.
- Unfairness in AI models can actually exacerbate unfairness in society itself through a kind of Barnesian performativity (the effect that using a model in practice makes a societal process more like its depiction by that model)

- Controlling such feedback may require a feedback-based strategy.

Social science-inspired governance principles

Current Approach	**Suggested Approach**
Self-Governance	Deliberative and inclusive governance with broad stakeholder involvement
Producer-Focused Governance	Ethics of co-responsibility, where producers and users assume shared responsibility
Static Governance	Governance built on a compositional calculus, paired with anticipation through technology foresight that specifically considers mobility and change
Dead-Reckoned Governance	Feedback-based approach with *ex post* surveillance, much like ongoing monitoring of drug safety