AI: The End of the World...of Work?

David A. Noever
PeopleTec, Inc.
Huntsville, AL
david.noever@peopletec.com

Matt Ciolino
PeopleTec, Inc.
Huntsville, AL
matt.ciolino@peopletec.com

Josh Kalin
PeopleTec, Inc.
Huntsville, AL
josh.kalin@peopletec.com

J. Wesley Regian
PeopleTec, Inc.
Huntsville, AL
wes.regian@peopletec.com

ABSTRACT

Specialized algorithms now routinely outperform human experts to greater than 90% accuracy and at greater persistent speeds for complex vision, language, and speech tests. Complex human cognition, however, hinges on chaining together many such specialized tasks into a larger workflow to mimic true perception. For better machine training and human job emulation, this research explores the question of stacking multiple tasks as building blocks or task chains. We examine two complex multi-stage human tasks: 1) language translation from either text, imagery (optical character recognition), or scene context; and 2) cursor-on-target for object detection. The first task, a universal translator, carries science fiction roots, but this test requires multiple cognitive steps to recognize an arbitrary chunk of characters whether in text, email, images, or audio. Given a requirement to translate any of the 7,000 human languages, we explore whether domain expertise matters when designing a universal translator, particularly when compared to having many specialized workers performing a single task. The second cognitive task corresponds to a geo-intelligence job: given any location on earth, find its overhead satellite view and context, count or identify each type of object, then finally generate a representative caption summarizing the scene. The original contributions of this work find that the errors of single tasks propagate multiplicatively as more specialized tasks get chained together. We further find that in total, many well-defined but small goal-oriented models can outperform human experts when presented with a difficult job as end-to-end pipelines within the scope of modern machine learning. We finally propose a series of key human tasks for future work that share similar features of multiple (>3) application programming interfaces (API). When called sequentially, job tasking in this way can generate flexible and often unexpected performance metrics with rapid learning curves.

ABOUT THE AUTHORS

David Noever has 27 years of research experience with NASA and the Department of Defense in machine learning and data mining. He received his Ph.D. from Oxford University, as a Rhodes Scholar, in theoretical physics and B.Sc. from Princeton University, summa cum laude, and Phi Beta Kappa. His primary research centers on machine learning, algorithms, and data mining for analytics, intelligence, and novel metric generation.

Matt Ciolino has research experience in deep learning and computer vision. He received his Bachelor's in Mechanical Engineering from Lehigh University. Matt is pursuing graduate study in computer vision and machine learning at Georgia Tech.

Josh Kalin is a physicist and data scientist focused on the intersections of robotics, data science, and machine learning. Josh received his Bachelor's in Physics, and Masters' in Mechanical Engineering from Iowa State University, and Computer Science from Georgia Tech.

J. Wesley Regian has 32 years of experience in cognitive performance modeling and knowledge-based software technology development, primarily for military applications with AFRL, AFOSR, and DARPA. His work has supported over 50 fielded systems. He has published over 100 papers on intelligence analysis, human terrain modeling, knowledge representation, and multi-source intelligence fusion.

AI: The End of the World...of Work?

David A. Noever
PeopleTec, Inc.
Huntsville, AL
david.noever@peopletec.com

Matt Ciolino
PeopleTec, Inc.
Huntsville, AL
matt.ciolino@peopletec.com

Josh Kalin
PeopleTec, Inc.
Huntsville, AL
josh.kalin@peopletec.com

J. Wesley Regian
PeopleTec, Inc.
Huntsville, AL
wes.regian@peopletec.com

INTRODUCTION

The modern concepts of AI evoke many societal fears and hopes, whether signaling the AI-apocalyptic end of the human-dominated world or just the beneficent end of the world of dull, repetitive work (Bughin, et al. 2018). A paradox in AI research thus poses machine intelligence as both essential and simultaneously threatening to our species' identity and survival (Nowak, et al., 2018). As qualified experimentally here, a modest approach poses the simpler question, what exactly are some machine learning capabilities today, and where might its multiple novel capabilities be headed tomorrow? This question in particular explores what in robotics has been called, "Moravec's paradox" (Mitchell, 2021), which states that "It is comparatively easy to make computers exhibit adult level performance on intelligence tests or playing checkers, and difficult or impossible to give them the skills of a one-year-old when it comes to perception and mobility" (Moravec, 1988). The surprising outcome may be that machines are good at hard things and bad at easy human things. To make machines better at this hard-easy dichotomy, one proposal has repeatedly attracted research attention, namely that complex behavior can spontaneously arise from chaining together many smaller tasks. This inspiration in part was an impressive Microsoft Research demonstration (Rashid, 2012). The authors demonstrated the multi-stage task chain where a non-native speaker addressed a Chinese auditorium. Their language translation automatically performed English speech-to-text (recognition), followed by text to Chinese character translation, then (logogram) text-to-speech back again in spoken Chinese Mandarin (Howard, 2014). This 4-step chain of machine learning models ran in real-time, accomplishing a hard task for humans but in an automated way (Rashid, 2012). Such task chains of multiple APIs seem to hold considerable power to get the ML field into more general learning and less specialized benchmarks on narrow tasks.

Figure 1. Example AI/ML Task Chain to Swap Video Backgrounds. *Built by authors using a single click to segment the foreground object of interest using RunwayML tools, (2021).*

In this context, the research searches (mainly by example) to find out whether complex perception and language skills might consist of simpler building blocks, or chained atoms, which combine to create or improve valuable tasks (Horn, et al., 2017). In Figure 1, we illustrate a three-stage video processing task, in which an algorithm first recognizes the human in one frame, then removes the snowy background, substitutes a green screen, and finally enables any background swap such as desert environments. All stages of this automated demo occur in an task chain across all subsequent video frames from a single frame.

Figure 2. Object Detection ML Layered on Map API

Figure 2 shows a second example task chain that overlays an airplane detection algorithm on top of an arbitrary location using the Google Map API with the user selecting latitude-longitude in one of our cloud-hosted web applications. While only the plane detection is machine learning in this example, the chaining of two specialized APIs together accomplishes a dynamic meta-task for the reconnaissance analysts.

The organization of the paper includes a survey of multiple application programming interfaces (APIs) (Hargil, et al. 2010) and then chains together the outputs of one machine learning specialization to serve as the inputs to another, and so forth. Horn, et al., (2017) refer to this instructional approach as cognitive task analysis (CTA) which elicits building blocks or hierarchies of skill chains for games and increasingly complex tasks. In total for this research, we will train 56 deep neural networks (DNN) and 5 generative adversarial networks (GANs). We apply these algorithms to complete 15 novel tasks by combining specialized subtasks to accomplish a larger meta- or super-task. By today's standards, this army of specially trained networks is not that overly ambitious but illustrates the field's rapid advancement and surveys possible near-term directions. Where noteworthy, the research identifies new capabilities derived from task chaining that reach beyond their initial specialization and singular API uses. A recent industry survey of over 13,000 respondents identified that 84.5% say that such "APIs are playing a significant role in their [digital transformation] initiatives." (Postman, 2020).

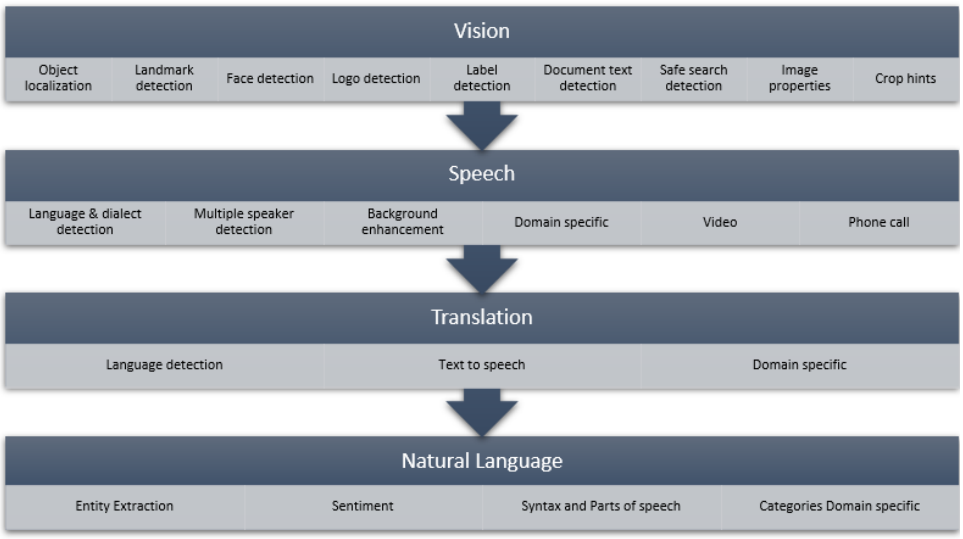

Figure 3. Task Chains for ISR Exploitation

With Figure 3 as our guide, we examine complex multi-stage human tasks that both need expertise and feature modular inputs and outputs. Similar to APIs, these tasks combine the output of the first model (or program) as the input to multiple downstream models.

METHODS

The research approach focuses on task chaining in both a serial and parallel way. The serial task involves steps in building a more complex natural language understanding where for instance the output of one task (text) is fed as the input to the next task (speech). For specificity, we consider 1) language translation from either text, imagery (optical character recognition), or scene context; and 2) cursor-on-target for object detection.

Language Translation and Speech-to-Text

To translate the estimated 7,111 unique languages, the combinatorial explosion of training data for bilingual sentence phrases exceeds known storage and computational limits. Handling the 10 most used languages alone requires 945 unique bilingual pairs for training sets and translation models. In February 2020, Google's Translation API supported 5 new languages covering about 75 million speakers, but also raising their total translatable language count to 108 and thus greater than 1086 bilingual pairs. For this task chain, we avoid the use of the public API interfaces like Bing and Google Translate services and illustrate how technologically accessible it is to build a universal, but personally customized translator. We then chain together multiple translators where the output from one language can become the input to another one, including closing the task loop by a full circle that returns to the originating language with an often altered result (Noever, et al. 2021). To demonstrate how these translators work, we rely heavily on training a neural network model first from scratch using only paired sentences in two different languages. We use a long short-term memory (LSTM), recurrent neural network (RNN) model as previously described in detail (Gers, et al., 1999; Noever, et al. 2021). The LSTM builds long-term associations or language context in a manageable set of neuron-like gates by over-weighting essential connections (like the closely related semantic terms, "lion-cat") in a process called "attention". Tuned for the key task of learning to translate, the LSTM slowly forgets over time ("short-term memory"). The successful text translator thus should generalize to previously unseen phrases and word combinations without memorizing a given set of training sequences. This task applies an LSTM network to language pairs in

multiple examples ranging from Hungarian to Japanese. The LSTM model trains its multiple layers, word embeddings, as more than 2.66 million tunable parameters. The bilingual pairs vary in length, but an average of 20,000 phrases includes about 110,000 words or the average scale of a 300 page novel as training data. Bilingual sentence pairs provide the core training data for Turkish, Hungarian, Dutch, German, and all other models. More than 81 phrase pairs of varying length have been published with examples from the Tatoeba Project (2020), as part of the Anki (2020) language flashcards software. The input for training is tab-delimited bilingual sentence pairs, with a third column attributing the reference for the (human) translated cases. In this crowd-sourced collection, the largest parallel training example for English sentences totals over 150,000 phrases and pairs that prompt input with a foreign language output for one single corpus.

Object Recognition Example with Image Captioning

To illustrate a bridge between vision and the previous language tasks, a second demonstration illustrates how an algorithm can caption or describe a satellite image as a relatively complete narrative. This ability to translate raw pixels to full sentences can prove useful for generating image search engines, the semantic grouping of objects in a hierarchy, or particularly for overhead imagery, shorten download delays by reducing image bandwidth into short text bursts (Noever, et al. 2020). Previous work (Cheng, et. al, 2017) has reviewed the challenges in classifying and captioning overhead imagery, noting particularly the lack of scene diversity (e.g. number of image classes). Various captioning benchmark efforts have appeared, some of which have offered a specialized earth-observation retrieval system based on the content of satellite imagery (Lu, et al. 2017) or have assembled a combined image and text dataset called the Remote Sensing Image Captioning Dataset (RSICD). The initial RSICD dataset (Lu, et al., 2017) consists of 10,921 satellite images broadly grouped into 30 characteristic scenes. RSICD particularly highlights land uses such as residential, urban, and agricultural classes. The RSICD authors follow a similar captioning format by providing five different sentences describing each image, Each satellite image is a small overhead chip, 224 x 224 pixels, collected at varying resolutions and sourced from overhead mapping services such as Google or Baidu. Thus, in total, the text portion of RSICD includes more than 50,000 sentences, 239,765 words, and the book equivalent of a 2100-page manuscript describing the 30 scene classes. Our approach to the image model highlights mainly the size and architecture of each approach within traditional transfer learning, where the model is pre-trained except for fine-tuning on the final object classification layers. We initially applied VGG-16 and VGG-19 (Visual Geometry Group, Oxford), which stands out among other image classification networks because of their architectural simplicity (Simonyan, et al., 2014). Only two building blocks are needed for VGG models, a 3x3 convolution and 2x2 pooling layer, throughout the entire network. Our research has tested seven new RNN-LSTM transfer-learning models for satellite image captioning (Noever, et al., 2020).

RESULTS

Language Translation and Speech-to-Text

The experimental results evaluate the potential for LSTM architectures to capture enough sequence order to build lightweight language translators using only bilingual phrase pairs (<20,000 pairs). Using this approach (Figure 4), we show the LSTM architecture compares favorably for some popular languages like Italian and outperforms large public translation services. We demonstrate that the bilingual pairs work in both directions, with English to Italian and vice versa. It is worth noting that the LSTM works reversibly with the bilingual pairs and requires no significant architectural modification to learn Russian-to-English for example versus English-to-Russian.

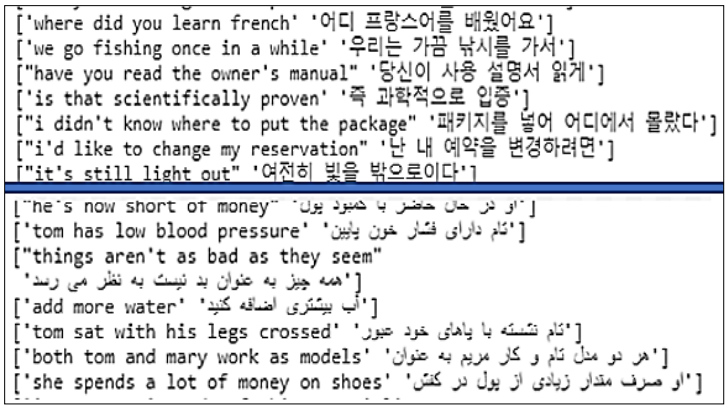

Figure 4. Example output test translation sequences in Korean (top) and Persian (bottom).

Object recognition example with Image Captioning

As shown in Figure 5 (B) for automated ISR tasks, our image-caption pairs can effectively describe the relations between overhead imagery and render new image repositories as searchable or discoverable datasets (Noever, et al, 2020). The algorithm generates the description (B) for the image of Korean factories as "Many buildings and some green trees in an industrial area." If this image proved uninteresting to the ISR analyst, rather than downloading many megabytes of satellite imagery, the analyst may deprioritize its availability using fewer than 50 text characters. Even in cases where sufficient ground stations could receive such tactical imagery, the text-only bandwidth savings typically exceeds a thousand-fold compared to full panchromatic transmission and analysis. To make the narrative richness of descriptions quantitative, we report (Noever, et al., 2020) that the best trained captioned model exceeds 0.65 in the BLEU (bilingual equivalency, Brownlee, 2019), which scores as fluent or narratively proficient compared to human labelers if the scores reach above 0.5.

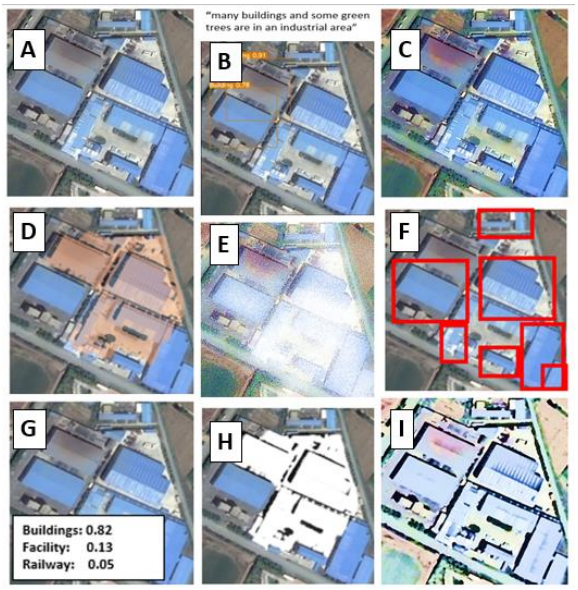

Figure 5. Automated ISR Tasks. *A. Original Satellite Image. B. Describe and Search Objects. C. Super-resolution. D. Decompose Materials. E. Remove Cloud Cover. F. Detect and Count. G. Classify Objects. H. Change Detection. I. Convert to Road Map.*

DISCUSSION

Why would be the implications if machine learning could achieve human level proficiency in language and vision in particular? The discussion section takes up this question after examining the previous two experimental examples in language and imagery exploitation.

Language Task Importance to Training Objectives

Militarily relevant foreign engagements naturally require language skills, a demand that has rapidly increased in post-9/11 operations. In the foreword to the Army's most recent doctrinal publication involving Special Forces (SOF), ADP 3-05, Special Operations, USASOC commanding general LTG Charles T. Cleveland writes that success in future conflicts requires "a profound understanding of foreign culture and fluency in local languages." (Walsh, 2014). Former SOCOM commander ADM William McRaven (2011) has frequently stressed the operational need for SOF personnel with "…languages, more cultural attunements, and regional expertise". The Defense Language Institute Foreign Language Center (DLIFLC, Monterey, CA) trains the military's interpreters. Depending on the language difficulty, the courses range from 36-64 weeks (Brown, 2017). Top of the list among the estimated 7,111 unique languages, Table 1 summarizes the hardest, easiest, and most in-demand languages for an English-speaking soldier to master (Taleninstituut Nederland, 2021; FluentU, 2021)). For training soldiers, foreign languages are grouped into four categories based on their difficulty (Walsh, 2014). Category I languages, such as French or Spanish, and Category II languages, such as Indonesian, are the easiest to learn, while Category III languages, such as Russian, and Category IV languages, such as Chinese, are the most difficult. In the spirit of competing machines against humans, the ready availability of mobile devices that access internet translation resources, one can imagine a future where the universal translator service can displace the human

Table 1. English-speaker Difficulty & Relevance for Translation Services

Hardest	Easiest	Civilian-Relevant	Militarily-Relevant
Mandarin Chinese	Norwegian	Arabic	Arabic
Arabic	Swedish	Basque	Mandarin Chinese
Japanese	Spanish	Mandarin Chinese	Pashto
Hungarian	Dutch	English	Farsi (Persian)
Korean	Portuguese	French	Russian
Finnish	Indonesian	German	Korean
Basque	Italian	Hindustani	Dari
Navajo	French	Italian	Urdu
Icelandic	Swahili	Japanese	

tutor, either through continuous learning or through augmenting human speech across multiple languages (Worsham, et al, 2020).

Object Recognition Task Importance to Training Objectives

The second cognitive task experimentally explored corresponds to a geo-intelligence job: given any location on earth, find its overhead satellite view and context, count or identify each type of object, then finally generate a representative caption summarizing the scene. As military sensors increasingly embed smarter pre- and post-processing to their standard operations, the task of an Intelligence, Surveillance, and Reconnaissance (ISR) Operator grow more efficient (Neisser, et al, 1963). As soldiers collect information about the adversary's strengths, movements, and activities, the individual training required extends beyond the 8.5 weeks of basic training, to include for Airborne Operators 2 weeks at Lackland Air Force Base for Aircrew Fundamentals Course, followed by 3 weeks for the Intelligence Fundamentals Course at Goodfellow Air Force Base. One main goal of this specialization follows from the need not only to participate in tactical mission planning, but also to understand how to operate, evaluate, and manage visual and signal processing systems. We examine this human skill within the context of recreating the human ISR analyst in a machine using algorithmic learning by examples. Mathur (2017) described a host of useful applications to generating text captions from images, including assisting disabled readers, enriching tags for social media searches, teaching children to read, and indexing the internet's image libraries. For military tasks specifically, Das et al. (2019) suggested that a pilot might speak a command, which when translated using speech-to-text, then could pull up various images for weather patterns, ISR recognition tasks, and radar on demand for hands-free comparisons with what the pilot sees in real-time. As shown in Figure 5, our research focuses on aiding the download problem for small satellites by tagging images of interest and downloading all low-bandwidth captions along with on-board localization of areas of interest.

CONCLUSIONS AND FUTURE WORK

Successful Task Chains

Figure 6. Task Chains for Example Image and Text Exploitation. *A. Tradeshow Image (Long, 2018). B. Facial Recognition. C. Chinese Text Extraction. D. Entity Extraction and Geo-location. E. Natural Language Processing. F. Landmark Detection.*

To illustrate task chains as more than the sum of their parts, we select a single image exploitation task (Long, 2018) in which a defense analyst breaks down a tradeshow exposition poster. The poster is shown in Figure 6 originally looks like a typical trifold poster with foreign language descriptions and a scaled rocket model (A). The task chain for the analyst is to employ only machine learning APIs and log everything that appears noteworthy from the image. We combine the Google Machine Learning APIs (see Figure 3 example list) to chain vision, speech, translation, and natural language tasks into a coherent narrative that arises from an entirely automated ML process. First, the **Vision API** yields a primary object recognized as "Aircraft" with 91% confidence. Thirty-three other labels are found in the same image with different confidence levels ranging from "sky" to "takeoff". The overall scene theme can be classified as appropriate for safe searches, e.g. not "medical, violent, spoof, racy or adult." Some noteworthy things the image does not flag include any particular landmarks (or geographic coordinates), faces, logos, or a single object to classify

in the scene. The key properties identified for the main "scaled rocket model", show its potential camouflage color as silver-grey [RGB(196, 197, 198)], which may prove useful to the analyst in recognizing the dominant metallic frame in future satellite imagery. The composite **Vision API** does find and extract (logogram) text in the picture, then correctly identifies the language as "Chinese" (C). When that foreign text output serves as the input to a **Translation API**, we find the transcoded pronunciation for the poster's banner as "*Jìn kōngjiān gāo chāoshēngsù tōngyòng shìfēi píngtái*", which the analyst can capture as a spoken audio file for later analysis using the **Speech API**. The English title identifies the poster's main point, which is to announce to the world that China has tested hypersonic aircraft, the Lingyun-1, a "*Near-space hypersonic universal test flight platform."* A perpendicular view of the same scene correctly finds a person, localizes the face (B), identifies the absence of blurring or headwear, then assesses the emotional effect from joy to anger. The **Vision API** identifies more foreign text from the close-up view of the poster's captions, which when passed through the **Translation** and **Natural Language APIs,** can extract key linguistic features including the launch date and place, the manufacturer, and assistance from particular published scientific references (including the Taiwanese Department of Math and Physics) (E). Using the launchpad information and date, the **Map API** can yield relevant overhead imagery around the same time (D). Finally using the initially published information about this trade show (Long, 2018), the analyst can identify that a military museum (F) has openly exhibited the Lingyun-1, which the **Vision API** correctly localizes by name and location from an unlabeled image using its ability to detect public landmarks. By task chaining multiple machine learning algorithms and a single trade show photo, the AI-augmented analyst automates in seconds what otherwise might languish in archives that take weeks to unravel and process to a report.

Cautionary Notes on Task Complexity and Adversarial Attacks

Deep learning models are well-understood as begin vulnerable to various adversarial attacks if slight modifications on the data or model parameters (Noever, 2021) reveal an incorrect decision boundary (Tramer, et al. 2017). We note that chained tasks may pose multiple entry points for such attacks, such that the global mission of vision or language gets side-tracked using any number of small modifications in the chain. Although propagating errors of this type are only beginning to be investigated (Noever, 2021), their consequences are likely more dramatic. As complexity increases, so do access points or vulnerabilities in a classic example of single-point failures in the chain (Cook, 1998). However, such complex systems in the real world typically multiply defensive layers in response to the inherent challenges of multi-step operations. Knight (2012) has noted a number of flaws remaining in existing speech recognition chains compared to human speech-to-text services, such as the rate of speech (e.g. fast is harder), the lack of background sounds (e.g. noise is harder), American white male speakers (e.g. ethnic women are harder), and professional domains or jargon (e.g. specialization is harder).

To illustrate another type of cumulative error, we apply APIs in a simple task chain for language translators. Figure 7 shows the children's telephone game or Chinese whispers, (Kontogiorgos, et al. 2020) spread through next-generation neural translation technology (Google Translate, 2021). With a starting military phrase, "On the frontline", we illustrate how a serial API chain can generate plausible translations when the algorithms perform single translation steps to the seven hardest languages from Table 1 (e.g. English-to-Arabic-to-English, which yields the correct roundtrip as "On the front lines"). The equivalent parallel chain, however, that stacks these API calls such that the output in one language becomes the input to an unrelated one, such that errors get magnified if one language step proves inferior (English-to-Arabic + Arabic-to-Chinese, etc.). The last column illustrates that after only three language steps, the phrase gets shortened to "In front" and degenerates like the telephone game into "in return" (Farsi), "To trade for" (Russian), and finally back to English as "the trade". In most cases for this error propagation (Wu, et al., 2018), the language difficulty magnifies the effect of repeated API calls.

Start Language	Serial API	Parallel API
English	***On the frontlines***	***On the frontlines***
Arabic	على الخطوط الأمامية	On the front lines
Mandarin Chinese	在前线	In front
Pashto	مخې ته	in front
Farsi (Persian)	در مقابل	in return
Russian	в обмен	in trade for
Korean	대한 무역	Trade for
Urdu	کے لئے تجارت	The trade
English	***The trade***	***The trade***

Figure 7. Telephone game with repeating translation through serial and parallel API cycles

The End of the World...of Work?

As a final demonstration of task chaining, we map the original research inspiration to specific multi-modal audio transformations. We describe a real-world 2019 cyber-attack, where a CEO's voice was cloned using machine learning, then replayed over the phone to the company's financial officer (Stupp, 2019). Recognizing the voice, a financial transaction was initiated and the company lost a quarter million dollars. One experimental version of this voice driven cyber-attack is shown in Figure 8. The illustration of all these steps is outlined in audio files that can be sampled and replayed online (Noever, 2021).

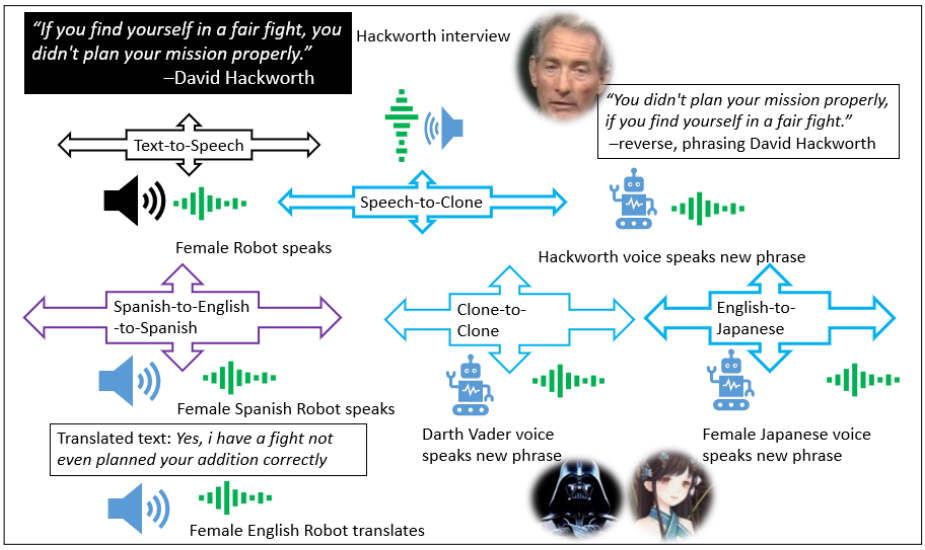

Figure 8. Audio task chains to demonstrate complex behaviors

The first stage of the cyber-attack is to convert English text to robotic speech using AI-driven Text-to-Speech (TTS) or speech recognition (Shen, et al., 2018; Prenger, et al., 2019). The output is shown as a female robotic voice in an audio (*.wav) file. Using this first task's output, we perform multiple chained transformations, including cloning another voice to speak various versions of the same voice (Jia, et al. 2018). Among our clone choices, we demonstrate the original speaker of the quote (the military journalist, David Hackworth) using a kind of style transfer in audio (Hackworth, 1989). This cloning now requires less than 10 seconds of recorded speech, which was captured from a cable interview on an entirely different subject using no overlapping words or phrases. Once we clone Hackworth's voice, we can record it speaking anything we type on the keyboard. To demonstrate this, we reverse the phrasing to make the sentence reordered. This cloning process can continue, which we show by transforming the Hackworth clone into Darth Vader speaking the same voice but trained on unrelated Star Wars phrases. This cycle demonstrates a complex task chain centered on the TTS (speech generation from text) to STT (speech recognition and translation to text).

A secondary task involves translation as the intermediaries shown in Fig 10. To complete an English-to-Spanish, and then back to English translation, we similarly go through a middle step that generates text from audio, then translates that text to another language. In our example, the voices are unchanged (female robot) but the text gets mangled in order and meaning when traveling full circle. Repeating this cycle across more than one language generates gibberish eventually, much like other artifacts of bad word substitutions or translation artifacts (Ciolino, et al., 2021).

This problem of compounding errors occurs not only in the text version, but we demonstrate it in cloned audio as well. First, we translate from English to Japanese text (Google Translate, 2021), then ask for the translation to be read thus creating an audio file purely in Japanese (female robotic voice). We then clone that translator to generate the speaking style in another language (English). The result is another round-trip through multiple transitions but remarkably can style transfer to a speaking rhythm and tone, but with zero training in English to generate this now English-speaking clone (Jia, et al. 2018).

While this convoluted path to a universal translator passes through a universal clone step, the result is a rather common military task (translation) but with an offensive element (voice-hacking, or vishing) in a cyber context. While not the end of translation work, it does represent a sophisticated attack rendered into a real threat by machine learning. If not the end of work, it could be imagined generating any fake message, biometric voice-print, or nuclear launch sequence, thus offering a rather odd way to arrive at the end of the world.

The Mechanical End of the Mundane?

Just as a human runner feels no need to sprint faster than a cheetah, most humans accept that a machine can calculate the square root of a large number better than any human, whether in this case, the term "better" refers to faster, more accurately, or just tirelessly (Mitchell, 2021). Humans concede the mathematical competition to an inexhaustible machine as useful to us but obscure to our survival and identity. Similarly, humans concede certain work categories to machines, particularly those jobs that qualify as dirty, dull, or dangerous. However, as observed for Moravec's paradox in perception and expression, modern machine learning begins to compete with many core beliefs of what a thinking human may hold dear and self-defining relative to other intelligent animals. If a machine can mimic the human processing of complex environmental cues such as sight and sound, then AI can staff a quality control job and serve as a factory inspector (Smith, et al., 2014). If a machine can mimic a tactile grasp and an adjustable grip, then AI can staff a warehouse job and work as a box packer. If a machine can mimic natural language, then AI can teach our children to read and write, or perhaps teach itself. The choices made here of calling out AI as mimicry, imitation, or an elaborate simulacrum are intentional. In the spirit of the original Turing test, if the intelligent agent can fool us with its skills, then the functional difference between mimicry and mastery recedes into philosophical realms. If we knew AI was ultimately just parroting us, then we still might not hesitate to delegate tasks to their repetitive roles. As listed in Table 2, specialized algorithms now routinely outperform human experts to greater than 90% accuracy and at greater persistent speeds for complex vision, language, and audio tests. For narrow job descriptions (like a help desk or call center) automation already dominates those jobs. The history of industrialization tends to reward specialists because of their efficiency or output quality. Complex human cognition, however, hinges on chaining together many such specialized tasks into a larger workflow. It's this concept of building more complex workflows that motivate this research. In this vein, we have identified particular concentrations where the powers of multiple algorithms are multiplicative in capabilities but not in compounding errors.

Exploitation Task	State of the Art Approaches
Entity Extraction	Stanford Named Entity Recognition (NER), Recurrent Neural Networks
Object Recognition	Tensorflow/ Keras / ResNet50, MobileNetV2, YOLO,
Scene Recognition	Semantic Segmentation, DeepLabV3, Deep Residual Networks
Time Series Prediction	Recurrent Neural Networks with Attention (RNN-A), ARIMA
Speech Recognition	Unidirectional Long-Term, Short-Term Memory (LSTM) Encoder
Time Series Anomaly Detection	Isolation Forests, DBScan, ARIMA, Recurrent Neural Nets (RNN)
Handwriting Recognition	Quasi-Recurrent Neural Networks (QRNN), Transformer
Facial Recognition	Deep face, Hidden Markov Models, Fuzzy Neural Nets
Human Action Recognition	Deep High-Resolution Representation Learning for Human Pose Estimation
Machine Translation	Convolutional Sequence to Sequence Learning, Translation

Table 2. Tasks and Algorithms for Building Example API Chains

ACKNOWLEDGEMENTS

The authors would like to thank the PeopleTec Technical Fellows program for its encouragement and project assistance.

REFERENCES

Asadi, K., Misra, D., Kim, S., & Littman, M. L. (2019). Combating the compounding-error problem with a multi-step model. arXiv preprint arXiv:1905.13320.

Babbel Magazine Nine Easiest Languages for English Speakers to Learn, (accessed 2021), babbel.com

Brown, D., The Military and Intelligence Community Need Linguists. Do You Have the Language Skills to Help Them? (2017) https://news.clearancejobs.com/2017/08/28/defense-language-jobs/

Brownlee, J, Deep Learning for Natural Language Processing. 2019. v.1.4, https://machinelearningmastery.com

Bughin, J., Seong, J., Manyika, J., Chui, M., & Joshi, R. (2018). Notes from the AI frontier: Modeling the impact of AI on the world economy. McKinsey Global Institute.

Chen, Y., Lu, X., & Wang, S. (2020). Deep Cross-Modal Image–Voice Retrieval in Remote Sensing. IEEE Transactions on Geoscience and Remote Sensing, 58(10), 7049-7061

Ciolino, M., Noever, D., & Kalin, J. (2021). Multilingual Augmenter: The Model Chooses. arXiv preprint arXiv:2102.09708.

Cook, R. I. (1998). How complex systems fail. Cognitive Technologies Laboratory, University of Chicago. Chicago IL..

Das, S., Jain, L., & Das, A. (2018, July). Deep learning for military image captioning. In 2018 21st International Conference on Information Fusion (FUSION) (pp. 2165-2171). IEEE.

FluentU, Foreign Language Immersion Online, “Earn Street Cred by Learning 1 of the 16 Coolest Languages” (accessed 2021), https://www.fluentu.com/blog/cool-languages-to-learn/

Gers, Felix A., Jürgen Schmidhuber, and Fred Cummins. "Learning to forget: Continual prediction with LSTM." (1999): 850-855.

Google Translate, (2021), https://translate.google.com/

Hackworth, D. H. (2020). About face: The odyssey of an American warrior. Simon and Schuster. See CSPAN Booktalks for voice clone sample, https://www.c-span.org/video/?7378-1/about-face-odyssey-american-warrior

Hargil, P., & Kapuscinski, C. (2010, October). Application Explosion: What's the right business model?. In 2010 14th International Conference on Intelligence in Next Generation Networks (pp. 1-7). IEEE.

Horn, B., Cooper, S., & Deterding, S. (2017, October). Adapting cognitive task analysis to elicit the skill chain of a game. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play* (pp. 277-289).

Howard, J. (2014). The Wonderful and Terrifying Implications of Computers that Can Learn, TEDxBrussels, https://www.ted.com/talks/jeremy_howard_the_wonderful_and_terrifying_implications_of_computers_that_can_learn?language=en

Jia, Y., Zhang, Y., Weiss, R. J., Wang, Q., Shen, J., Ren, F., ... & Wu, Y. (2018). Transfer learning from speaker verification to multispeaker text-to-speech synthesis. arXiv preprint arXiv:1806.04558.

Kalin, J., Noever, D., Ciolino, M., Hambrick, D., & Dozier, G. (2021, April). Automating defense against adversarial attacks: discovery of vulnerabilities and application of multi-INT imagery to protect deployed models. In Disruptive Technologies in Information Sciences V (Vol. 11751, p. 117510I). International Society for Optics and Photonics.

Knight, M. (2012). Speech Recognition Part I, http://blog.stenoknight.com/2012/05/cart-problem-solving-series-sitting.html

Kontogiorgos, D., Sibirtseva, E., & Gustafson, J. (2020, May). Chinese whispers: A multimodal dataset for embodied language grounding. In Proceedings of The 12th Language Resources and Evaluation Conference (pp. 743-749).

Long, D., (2018), China reveals Lingyun-1 hypersonic missile at National Science and Technology expo, The Defense Post, https://www.thedefensepost.com/2018/05/21/china-lingyun-1-hypersonic-missile-revealed/

Mathur, P., Gill, A., Yadav, A., Mishra, A., & Bansode, N. K. (2017, June). Camera2Caption: a real-time image caption generator. In 2017 International Conference on Computational Intelligence in Data Science (ICCIDS) (pp. 1-6). IEEE.

McCann, B., Keskar, N. S., Xiong, C., & Socher, R. (2018). The natural language decathlon: Multitask learning as question answering. arXiv preprint arXiv:1806.08730.

McRaven, W. Written Statement to the Senate, Advanced Policy Questions for Vice Admiral William H. McRaven, USN: Nominee for Commander, United States Special Operations Command, Senate Confirmation Hearing, 28 June 2011, 30.

Mitchell, M. (2021). Why AI is Harder Than We Think. arXiv preprint arXiv:2104.12871.

Moravec, H. (1988). Mind children: The future of robot and human intelligence. Harvard University Press.

Neisser, U., Novick, R., & Lazar, R. (1963). Searching for ten targets simultaneously. Perceptual and motor skills, 17(3), 955-961.

Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. Cognitive skills and their acquisition, 1(1981), 1-55.

Noever, D., Regian, J.W. (2019). Reinforcement Learning for Automated Textual Reasoning. In Proceedings of Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC). Orlando, FL, 2 Dec-6 Dec 2019.

Noever, D. Audio Samples for Machine Learning, unpublished, (2021) https://deeperbrain.com/demo/aud.html

Noever, D., Regian, W., Ciolino, M, Kalin, J., Hambrick, D. "Novel Scoring with Confusion Matrices for Satellite Image Captioning", 2020 Southern Data Science Conference, August 12-14 2020, Atlanta, GA

Noever, D., Kalin, J., Ciolino, M., Hambrick, D. and Dozier, G. "Local Translation Service for Neglected Languages", 2nd International Conference on Natural Language Processing, Information Retrieval and AI (NIAI 2021), 23-24 Jan 2021 & proceedings published by Computer Science Conference Proceedings in Computer Science & Information Technology (CS & IT) series

Nowak, A., Lukowicz, P., & Horodecki, P. (2018). Assessing Artificial Intelligence for Humanity: Will AI be the Our Biggest Ever Advance? or the Biggest Threat [Opinion]. IEEE Technology and Society Magazine, 37(4), 26-34.

Pramanik, S., Agrawal, P., & Hussain, A. (2019). Omninet: A unified architecture for multi-modal multi-task learning. *arXiv preprint arXiv:1907.07804*.

Prenger, R., Valle, R., & Catanzaro, B. (2019, May). Waveglow: A flow-based generative network for speech synthesis. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 3617-3621). IEEE.

Postman, 2020 State of the API Report, (2020), https://www.postman.com/state-of-api/#key-findings

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Rashid, R. (2012). Speech Recognition Breakthrough for the Spoken, Translated Word. https://www.youtube.com/watch?v=Nu-nlQqFCKg

RunwayML: Machine Learning for Creators, (2021), https://github.com/runwayml Original video source available here https://media.giphy.com/media/RVFkBglC4vkXbVxgJs/giphy.mp4

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., ... & Wu, Y. (2018, April). Natural TTS synthesis by conditioning Wavenet on MEL spectrogram predictions. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 4779-4783). IEEE.

SINTEF. (2013, May 22). Big Data, for better or worse: 90% of world's data generated over last two years. ScienceDaily. Retrieved June 13, 2021 from www.sciencedaily.com/releases/2013/05/130522085217.htm

Smith, A., & Anderson, J. (2014). AI, Robotics, and the Future of Jobs. Pew Research Center, 6, 51.

Stupp, C. (2019), Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case, Wall Street Journal, https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402

Sundermeyer, M., Ralf Schlüter, and Hermann Ney. (2012)"LSTM neural networks for language modeling." In Thirteenth Annual Conference of The International Speech Communication Association.

Taleninstituut Nederland, The Hardest Languages In The World To Learn, (2021), https://taleninstituut.nl/the-hardest-languages-in-the-world-to-learn/

Tatoeba Project, Accessed (2020), https://tatoeba.org/eng/sentences/index

Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., & McDaniel, P. (2017). Ensemble adversarial training: Attacks and defenses. arXiv preprint arXiv:1705.07204.

Walsh, S. (2014) Special Forces Language Training: What Would It Cost To Do It Right?, Small Wars Journal

Wang, X., Yan, H., Huo, C., Yu, J., & Pant, C. (2018, August). Enhancing Pix2Pix for remote sensing image classification. In 2018 24th International Conference on Pattern Recognition (ICPR) (pp. 2332-2336). IEEE.

Weiss, R. J., Chorowski, J., Jaitly, N., Wu, Y., & Chen, Z. (2017). Sequence-to-sequence models can directly translate foreign speech. arXiv preprint arXiv:1703.08581.

Worsham, J., & Kalita, J. (2020). Multi-task learning for natural language processing in the 2020s: Where are we going?. Pattern Recognition Letters, 136, 120-126.

Wu, L., Tan, X., He, D., Tian, F., Qin, T., Lai, J., & Liu, T. Y. (2018). Beyond error propagation in neural machine translation: Characteristics of language also matter. arXiv preprint arXiv:1809.00120.

Xie, L., Pan, P., Lu, Y., & Wang, S. (2014, November). A cross-modal multi-task learning framework for image annotation. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management* (pp. 431-440).

Zhao, X., Jia, Y., Li, A., Jiang, R., & Song, Y. (2020). Multi-source knowledge fusion: a survey. *World Wide Web*, *23*(4), 2567-2592.

Zou, Z. (2020). Castle in the Sky: Dynamic Sky Replacement and Harmonization in Videos. arXiv preprint arXiv:2010.11800.