

Bias in AI (machine learning) Ethics, Legality & Society

**CPSC 433: Artificial Intelligence
Fall 2024**

Jonathan Hudson, Ph.D.
Assistant Professor (Teaching)
Department of Computer Science
University of Calgary

August 8, 2024

Copyright © 2024

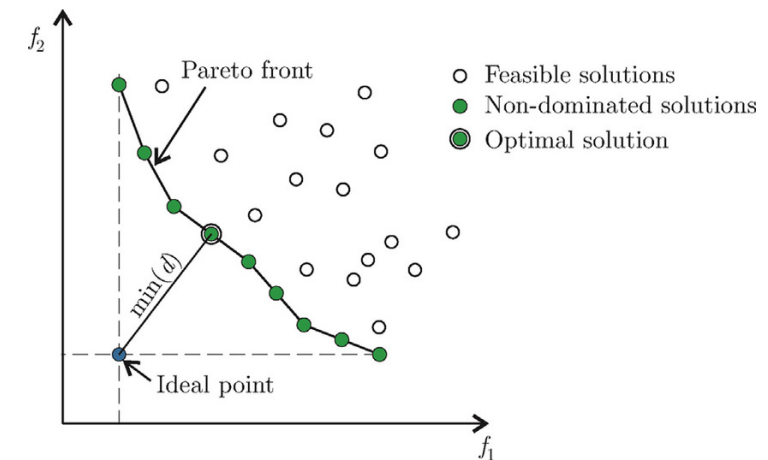


**UNIVERSITY OF
CALGARY**

Society

AI in society

- We have defined AI by rationality
 - Optimizing a 'utility function'
 - Similar to slippery slope to 'eugenics', government decides who gets to reproduce or exists (based on?)
- Issues
 - The real-world is not a numerical space
 - Many problems are non-singular values
 - Multi-modal
 - There are methods to optimize 'pareto front'
 - Benefitting 99% of people, does not make something 'fair'
 - Your value system is not universal
 - We live in a society, where balances must be negotiated
 - stakeholders, business vs individual vs community balances, etc.
 - We do have stakeholder, voting AI techniques



Indigenous and normative western AI

- “We don't have an AI ethics problem, we have an AI epistemology problem,” Jason Edward Lewis, Concordia University
 - epistemologies (or theories of knowledge)
- “a normative western approach that is favoured by current AI research is the assumption that the user is an individual, and the individual prioritizes their own well-being.”
 - Ex Machine Learning is designed to train for individual output correctness (individual)
- “this creates a blindness to vital aspects of human existence—such as trust, care and community—that are fundamental to how intelligence actually operates.”
 - “Collapsing intelligence down to a rational, goal-seeking, self-serving agent—in all of our cultures, that kind of person would be seen as selfish, foolish, not a good community member, and not intelligent.”

https://www.sshrc-crsh.gc.ca/funding-financement/nfrf-fnfr/stories-histoires/2023/inclusive_artificial_intelligence-intelligence_artificielle_inclusive-eng.aspx

Western AI Overview

- For AI (and Data rights) currently reasonably fair in English politics to place EU on the side of citizen's rights first, and the US on the side of business rights first
 - (see history of GPDR – General Data Protection Law) -> where American business regularly have to add protections for users to operate the same product in the EU
 - Ex. Think cookie protections, right to be forgotten, data hosting requirements
 - Canada (and nations like Australia) regulatory wise usually fall closer to US permissiveness due to economic pressures
 - UK is generally balance of need to fall closer to EU to cooperate but also alternate political pressures to distance themselves and try and be more like US

Non-Western AI Overview

- Outside of EU very few nations of explicit AI law (a few have AI addendums into prior law
 - ex. US Federal Aviation Administration (FAA) had purview of AI formally clarified in additional specifications
 - Much of this came from concerns over Boeing issues with economics, testing, and deployment of automated flight components in Boeing 737 Max crashes
- The other most influential nations are China/India where, like the English nations, law is very much influenced by politics and economics
 - China – (no existing direct AI law) differences in influences include ‘government approval list’ for areas of AI, leniency on copyright (business permissive has been prior history in courts)
 - India – (no existing direct AI law) differences in influences include prior evidence of lack of government structure to achieve oversight and regulation at scale withing country, many business agreements with international companies with differing principles (permissive preferred)

Legality

EU AI Act

- EU AI Act (March 2024 - final)
 - Rules based on risk of AI system
 - Unacceptable (Banned) - Cognitive behavioural manipulation of people or specific vulnerable groups, social scoring, biometric identification/categorisation of people, real-time biometric systems (ex. Facial recognition)
 - High risk
 - Cat 1 Safety products – toys, aviation, cars, medical, elevators (under safety laws)
 - Cat 2 Non-safety – infrastructure, education, employment, access to services/benefits, law enforcement, migration, legal interpretation (requires registration with EU)
 - Require assessment before market and during lifecycle, right of file complaints to authority
 - Transparency
 - Generative is not high risk but must comply with EU copyright law
 - Disclose AI generated, prevent from making illegal content (can this be done?), publish summaries of use of copyrighted works when training

Canada

- AIDA (Canada's Artificial Intelligence and Data Act)
 - As of 2024, Bill C-27, encompassing AIDA (second reading -> in House of Commons)
 - The AIDA proposes the following approach:
 - ensure that high-impact AI systems meet the same expectations with respect to safety and human rights to which Canadians are accustomed.
 - prohibit reckless and malicious uses of AI that cause serious harm to Canadians and their interests through the creation of new criminal law provisions.
 - plans to sync 'high-impact' with things with EU AI Act
 - High-impact - Evidence of risks of harm to health and safety, or a risk of adverse impact on human rights; The severity of potential harms; The scale of use; The nature of harms or adverse impacts that have already taken place; The extent to which for practical or legal reasons it is not reasonably possible to opt-out from that system; Imbalances of economic or social circumstances, or age of impacted persons; and The degree to which the risks are adequately regulated under another law.

<https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act-aida-companion-document>

Canada

- Of interest systems
 - Screening for employment (ex. video screening, resume filtering)
 - Biometric systems (ex. facial recognition)
 - Systems to influence behaviour at scale (ex. content algorithms)
 - Health and safety (ex. autonomous driving, triage systems)
- The AIDA addresses two types of adverse impacts associated with high-impact AI systems.
 - harms to individuals.
 - systemic bias in AI systems in a commercial context.
- Harm includes physical harm, psychological harm, damage to property, or economic loss to an individual.
- Harms may be experienced by individuals independently or may be experienced broadly across groups of individuals.
- Biased output occurs when discrimination falls under the Canadian Human Rights Act.

Canada

- Adverse differentiation could be considered justified if it is unavoidable in the context of real-world factors affecting a decision or recommendation.
- For example, individual income often correlates with the prohibited grounds, such as race and gender, but income is also relevant to decisions or recommendations related to credit.
- The challenge, in this instance, is to ensure that a system does not use proxies for race or gender as indicators of creditworthiness.
- For example, if the system amplifies the underlying correlation or produces unfair results for specific individuals based on the prohibited grounds, this would not be considered justified.

Australia

- No current AI law (some AI principles since 2019) – benefit, human-centred, fairness, privacy, reliability/safety, transparency/explainability, contestability, accountability)
- Regulators have declared personal information can't be used in public LLMs due to privacy risks (chat-gpt)
- Have 2 categories within AI
 - AI – engineered system that generates **predictive outputs** such as content, forecasts, recommendations or decisions for a given set of human-defined objectives or parameters without explicit programming (more like machine learning than just AI)
 - ADM (automated decision making) - refers to the application of automated systems in any part of the decision-making process (would include non machine learning AI that just uses rules)

US

- Like Canada/Australia, no formal AI law
- In US, companies are legally individuals, and lobbying has few to no limits
- The same reason nothing like GPDR exists in US is same reason any larger AI law will take a long time and have limited strength
- Much of AI law will be state dictated
 - Most useful to watch AI law in California and secondly New York, their population size and economic influence means their decisions are likely to create shadow national law
- President has passed executive orders which direct federal agency behaviours around things like algorithm discrimination, individual rights, agencies having AI officers
- Until 2019, most of lawmakers' attention around AI was absorbed by autonomous or self-driving vehicles and concerns about AI applications within the national security arena.

Some California AI Laws

- Privacy laws clarified as applying to generative AI outputs
- AI literacy in schools
- Healthcare must disclose generative AI usage, limits of health care automation to require supervision of AI usage
- Robocalls must disclose use of AI voices
- Child abuse images include those generated by AI
- ‘Nude’ deep fakes illegal to blackmail with, social media required to report ‘nude’ deep fakes
- AI—generated images require watermarks
- Election deep-fake laws
- Need permission to make AI replica of actors (alive or dead)
- Biometric Information Bill (data rights to biometric data)

Bias

Possible Sources for Bias in AI Systems

In Knowledge Representation:

- Missing knowledge (can't learn what's not there)
- Extra knowledge (important knowledge can look less important if smaller)
- Manipulated knowledge (incorrect data)
- Real-world knowledge that was already biased

In Knowledge Processing:

- Leaving out some transitions
- Wrong goal function in control
- Overall control heuristic if not enough search time

Leaving out knowledge

Since AI systems are knowledge-based, not providing all knowledge or not all possible examples for the system to learn the knowledge, obviously creates biases.

Bias source:

- AI system designer or the knowledge sources/experts

Counter measures:

- Use several (independent) knowledge sources
- Enable system to collect its own examples
- Test examples for “white spots” (gaps in system)
 - Ex. People with darker skin colour your algorithm fails to identify

Controlling accessible information

Manipulating the information an AI system can access obviously also can be used to create a bias

Bias source:

- People who can influence the environment of an AI system

Counter measures:

- Enable more information gathering using more “sensors”

Leaving out transitions

Transitions determine where a search can go. Leaving out some means that certain states cannot be reached and therefore certain solutions might become unreachable. Changing transitions naturally also can achieve that.

Bias source:

- System designer/programmer

Counter measures:

- Design reviews
- Code reviews
- Good testing

Using wrong control function

Evaluating the current state requires criteria and important criteria can simply be left out of the control function or the weights of the different criteria can be biased against them.

Bias source:

- System designer/programmer

Counter measures:

- Design reviews
- Code reviews
- Good testing

Search control heuristic has not enough time

Most search heuristics use knowledge to speed up the search for many search instances. And often tiebreakers! If the search is stopped, before the wanted end criterion is fulfilled, the solution is biased by the heuristic.

Bias source:

- AI developer
- Complexity of instance (→ the universe)

Counter measures:

- More computing power, more distribution
- More heuristics
- “pray”

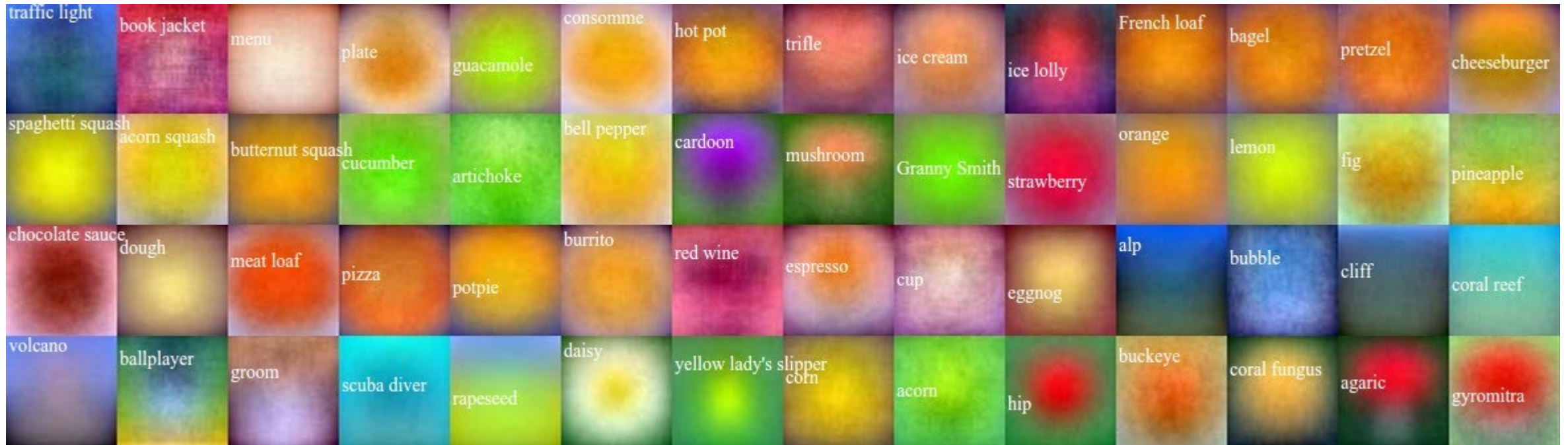


Neural Network Lesson

Linear Regression

Breaking Linear Regression

- Linear classifiers that takes every input pixel and maps to labels
- Take some food ones and back reverse to find image colours



Breaking Linear Regression

- For example, Granny Smith apples are green, so the linear classifier has positive weights in the green color channel and negative weights in blue and red channels, across all spatial positions. It is hence effectively counting the amount of green stuff in the middle.
- trick the Granny Smith classifier
 1. figure out which pixels it cares about being green the most
 2. tint those green
 3. profit!

Neural Networks are Logistic Regressions

- We are basically training a large function
- Finding its weights
- A fundamental struggle will always be the exploitability of this exact back-relationship of input and output

- Doesn't actually 'think' on an abstract conceptual level at this time
- We can find and reverse engineer mistakes based on trivial signals

Adversarial

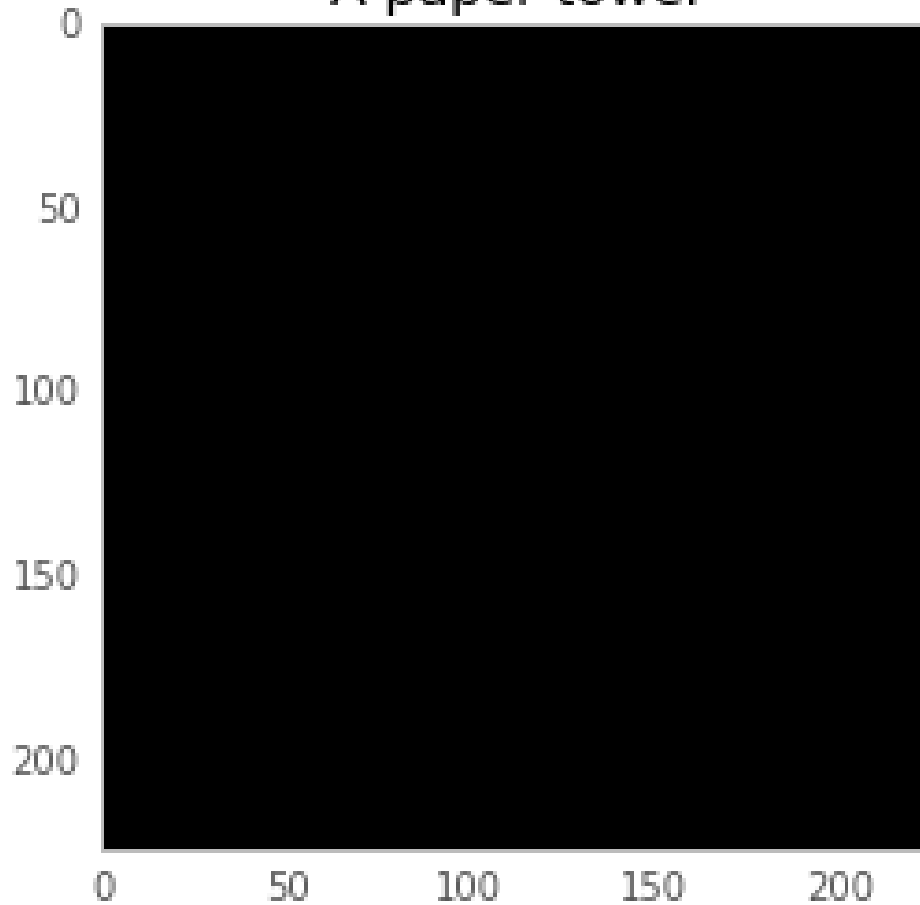
- We are usually nice to neural networks
- We feed it data like it has seen before and get back positive results
- Instead what if we are malicious and exploit how they work
- Main point (neural networks are basically very complicated functions which we can back solve and exploit)

Neural Network Lesson

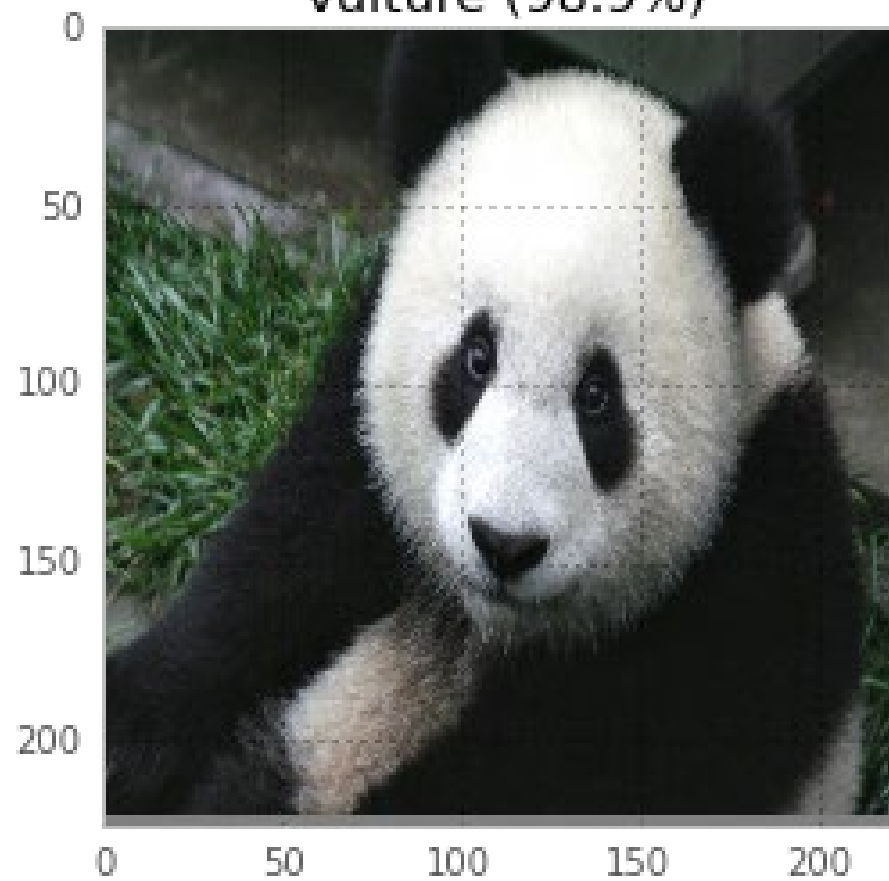
Adversarial Design

What!

A paper towel



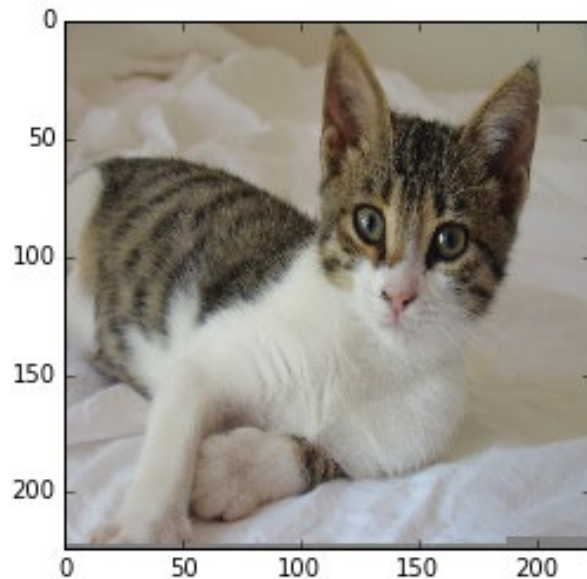
Vulture (98.9%)



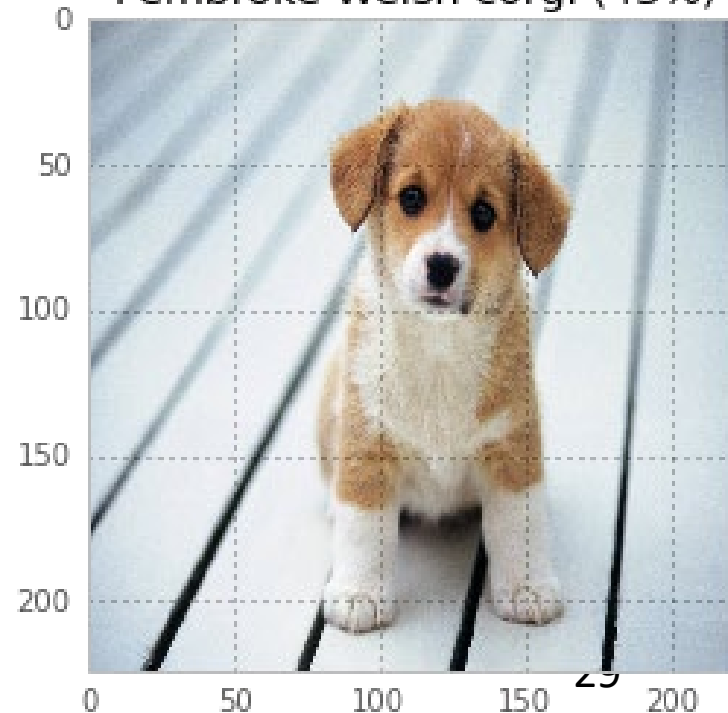
GoogLeNet

- These examples are GoogLeNet from the ImageNet competition
- So a large CNN, that should be (slightly) more robust against exploitation

class: 285
label: n02124075 Egyptian cat
certainty: 34.57%



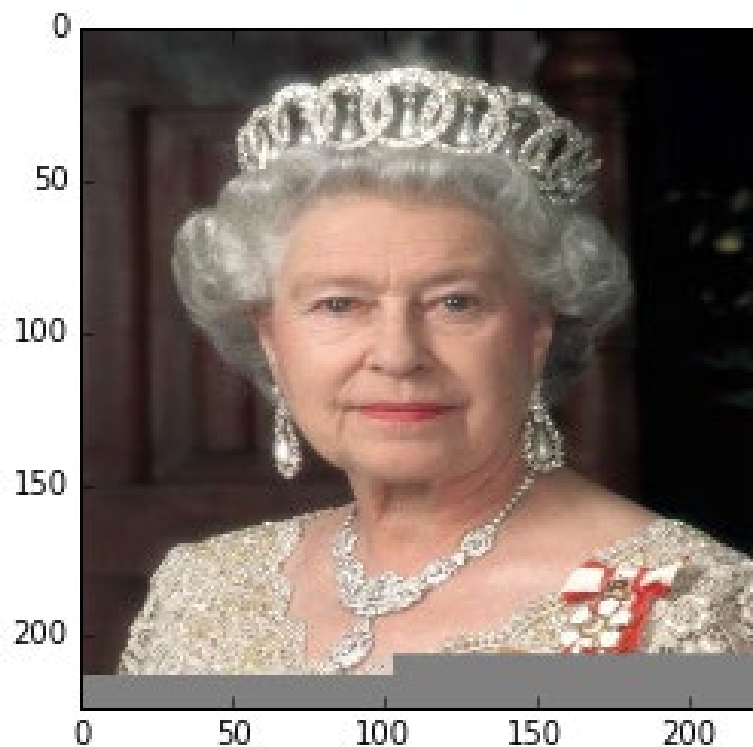
Pembroke Welsh corgi (43%)



GoogLeNet

- Even good CNNs can suck without our help

```
class: 793  
label: n04209133 shower cap  
certainty: 99.7%
```

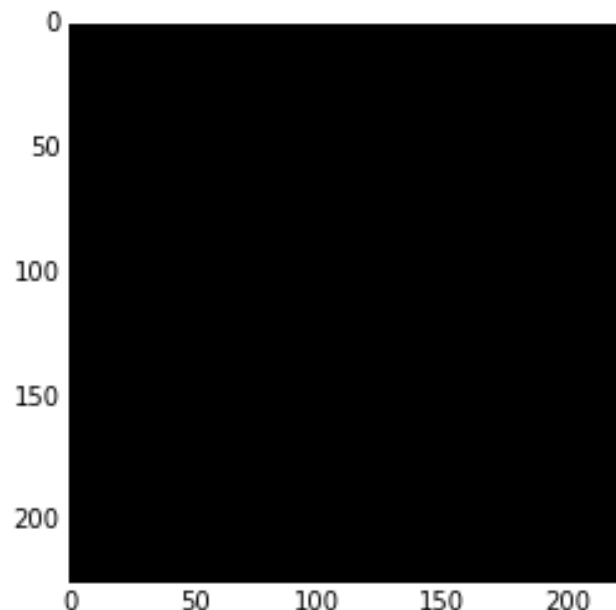


Let's dig around inside

- Make all black input
- Look at labels
- Even non-data has classification
- We are going to play with gradients

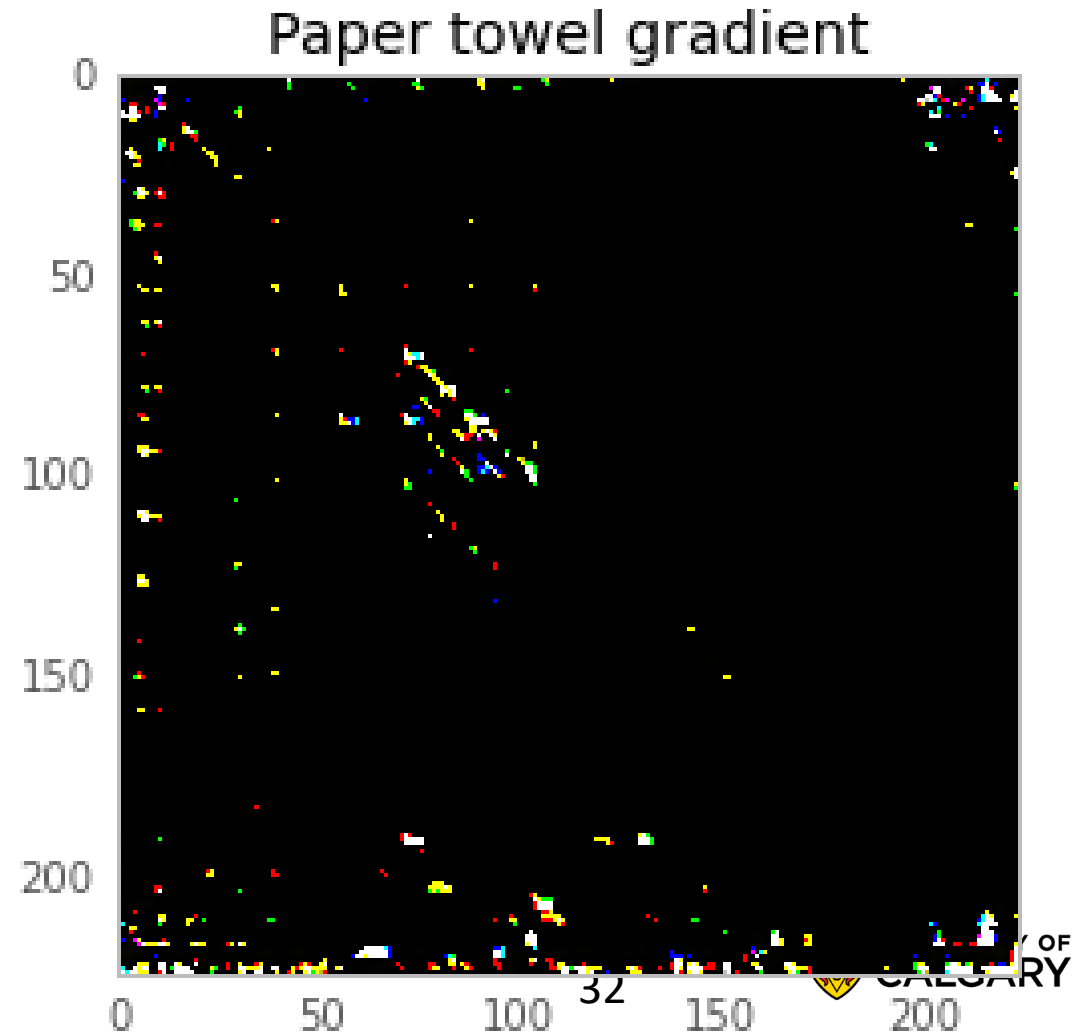
```
black = np.zeros_like(grad) * 255  
_ = predict(black, n_preds=5)
```

```
label: 885 (velvet), certainty: 27.38%  
label: 794 (shower curtain), certainty: 6.4%  
label: 911 (wool, woolen), certainty: 6.19%  
label: 700 (paper towel), certainty: 4.67%  
label: 904 (window screen), certainty: 4.39%
```



Reverse back-propagation

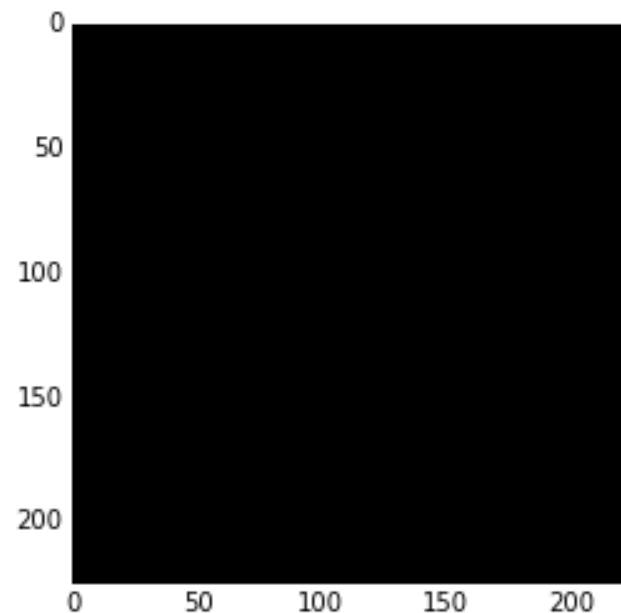
- Take paper towel as a label
- Set it to a full 1
- And back propagate the neurons



Reverse back-propagation

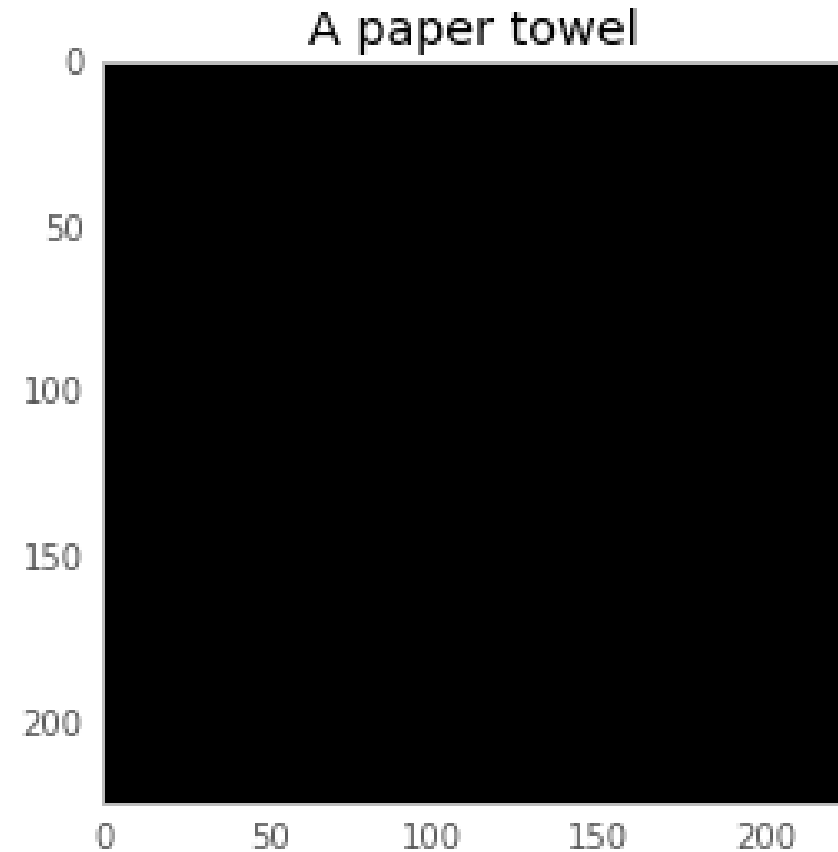
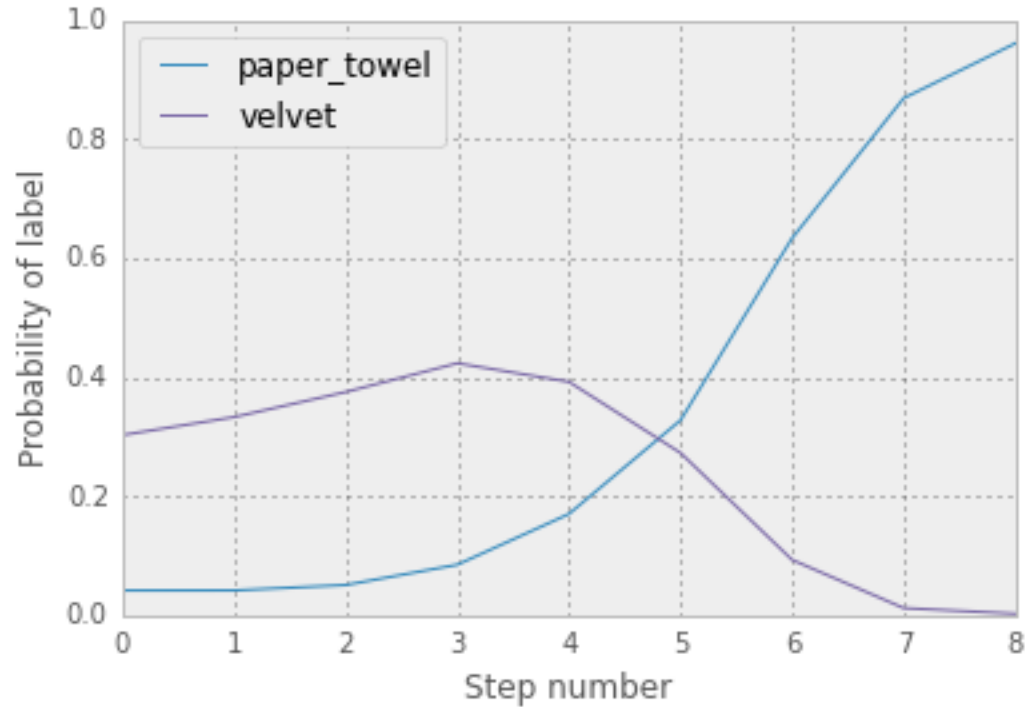
- We can see the garbage input ourselves
- So let's drop the ratio to 1/256
- We went from 4.67 to 16.03 %
- On something that still looks black to us

```
_ = predict(black + 0.9*delta, n_preds=5)  
label: 885 (velvet), certainty: 54.75%  
label: 700 (paper towel), certainty: 16.03%  
label: 911 (wool, woolen), certainty: 12.4%  
label: 533 (dishrag, dishcloth), certainty: 2.65%  
label: 794 (shower curtain), certainty: 2.11%
```



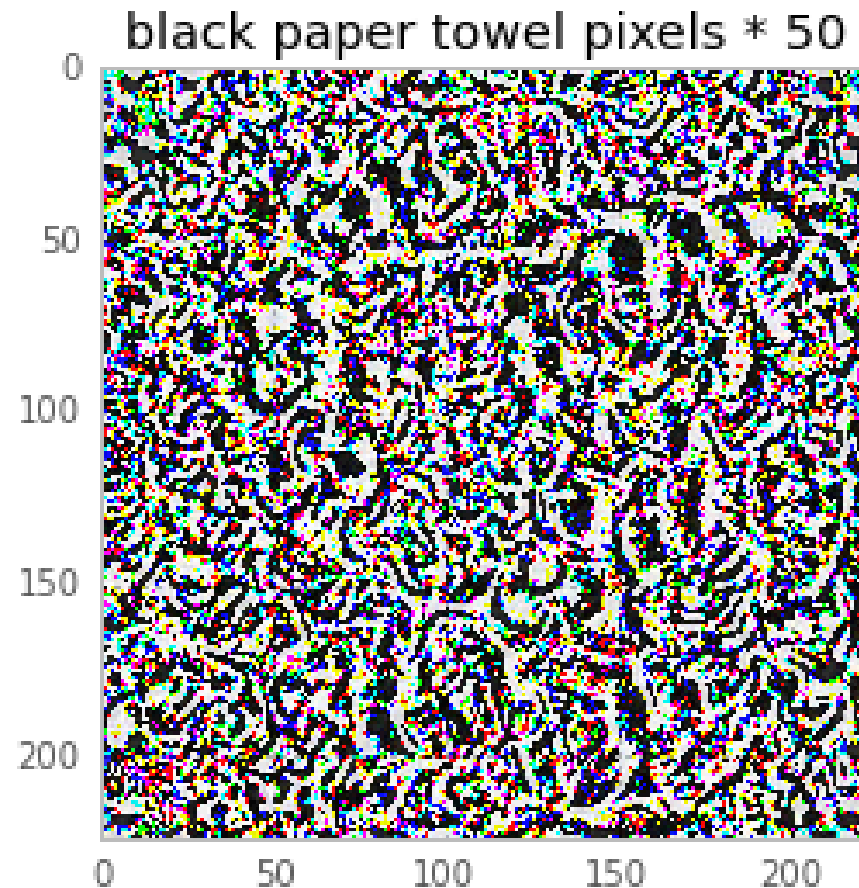
Reverse back-propagation

- Looping back propagation

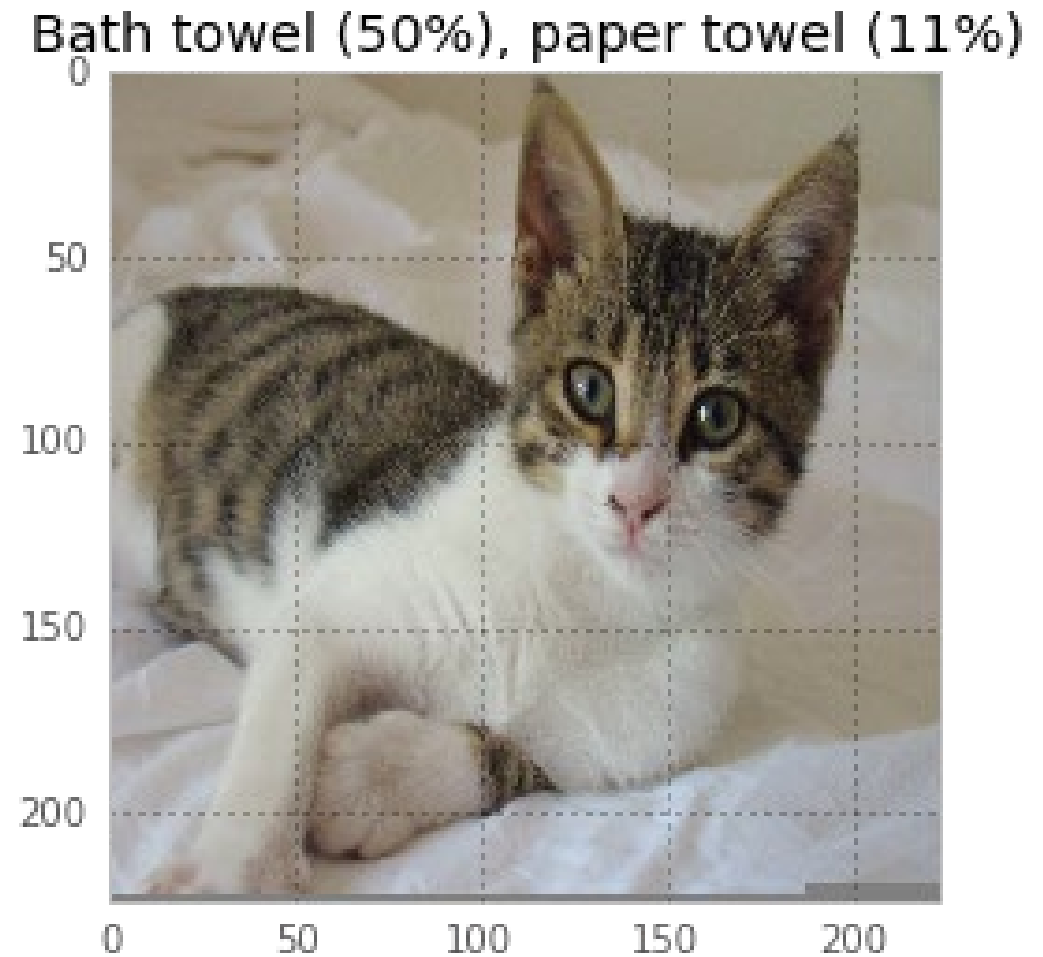
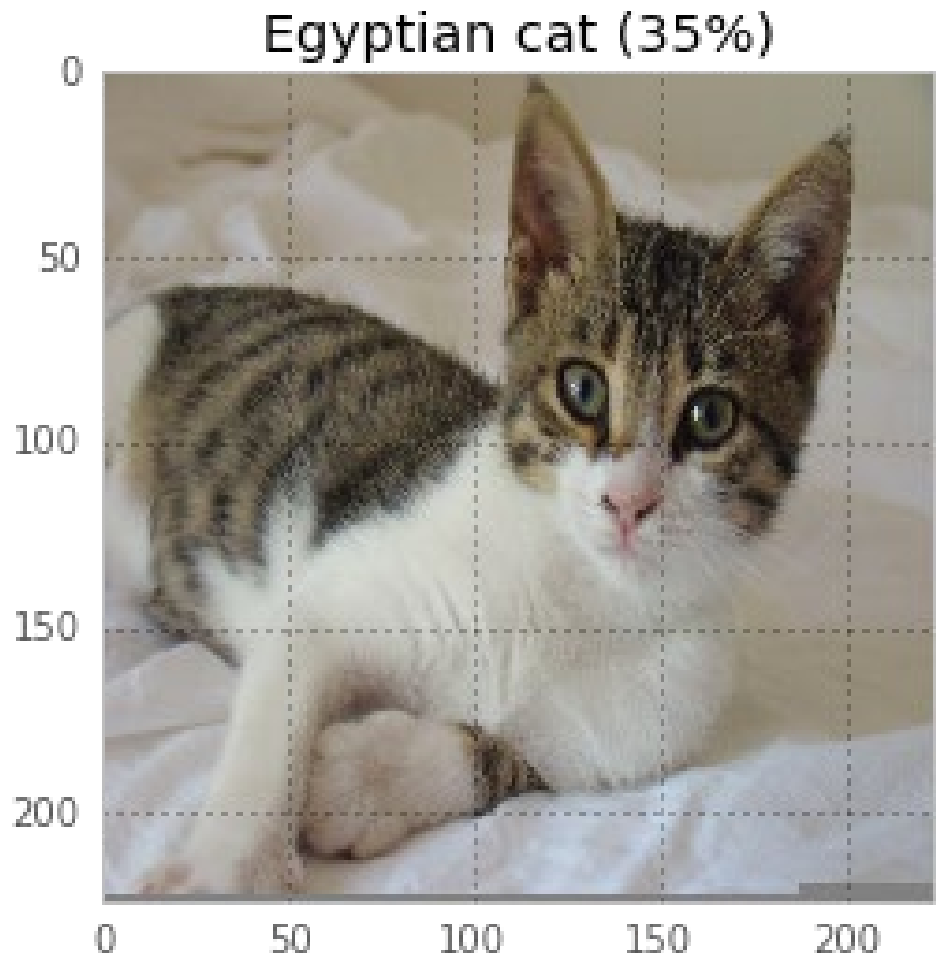


Reverse back-propagation

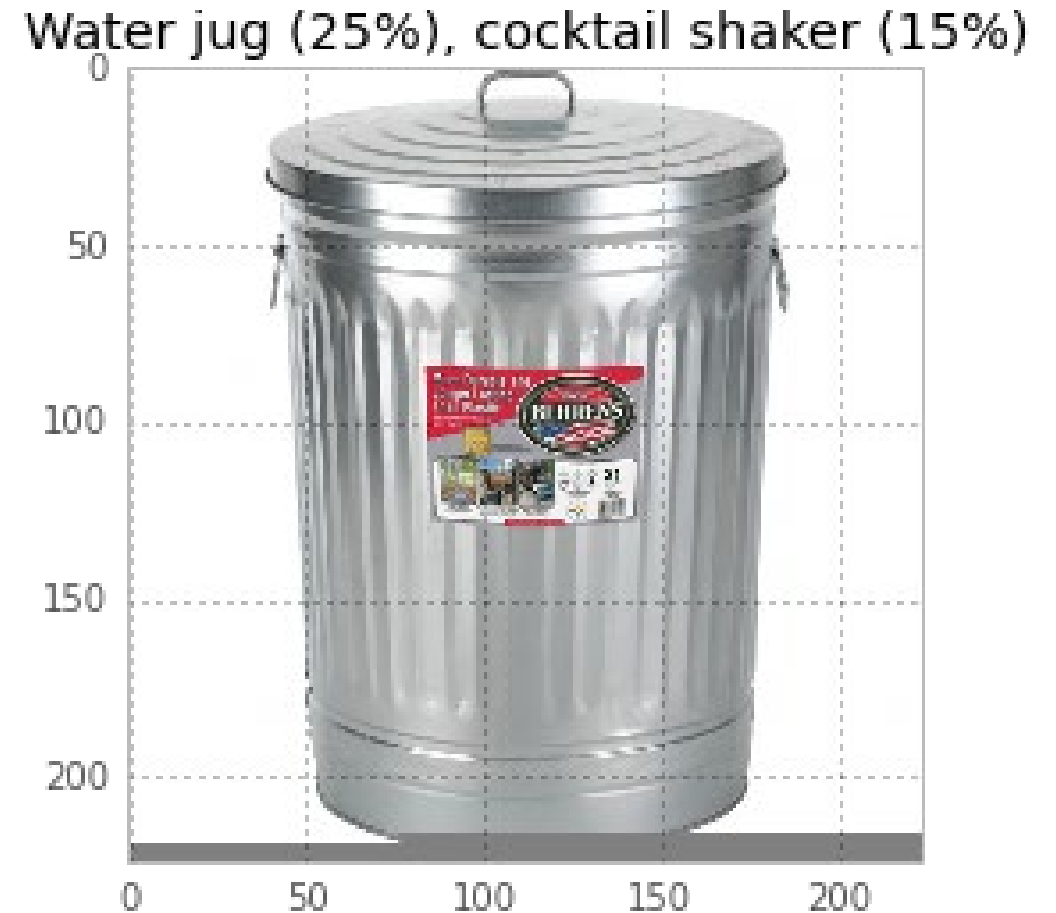
- Force the pixel values larger so we can see underlying structure



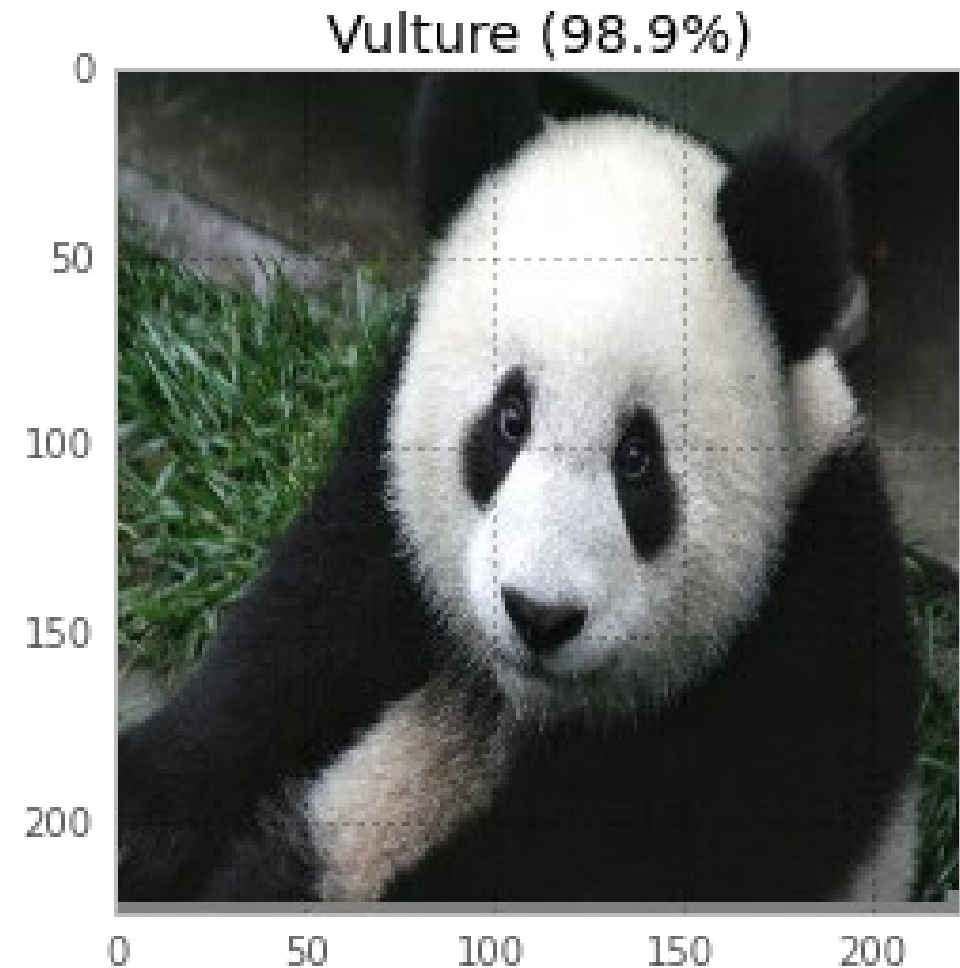
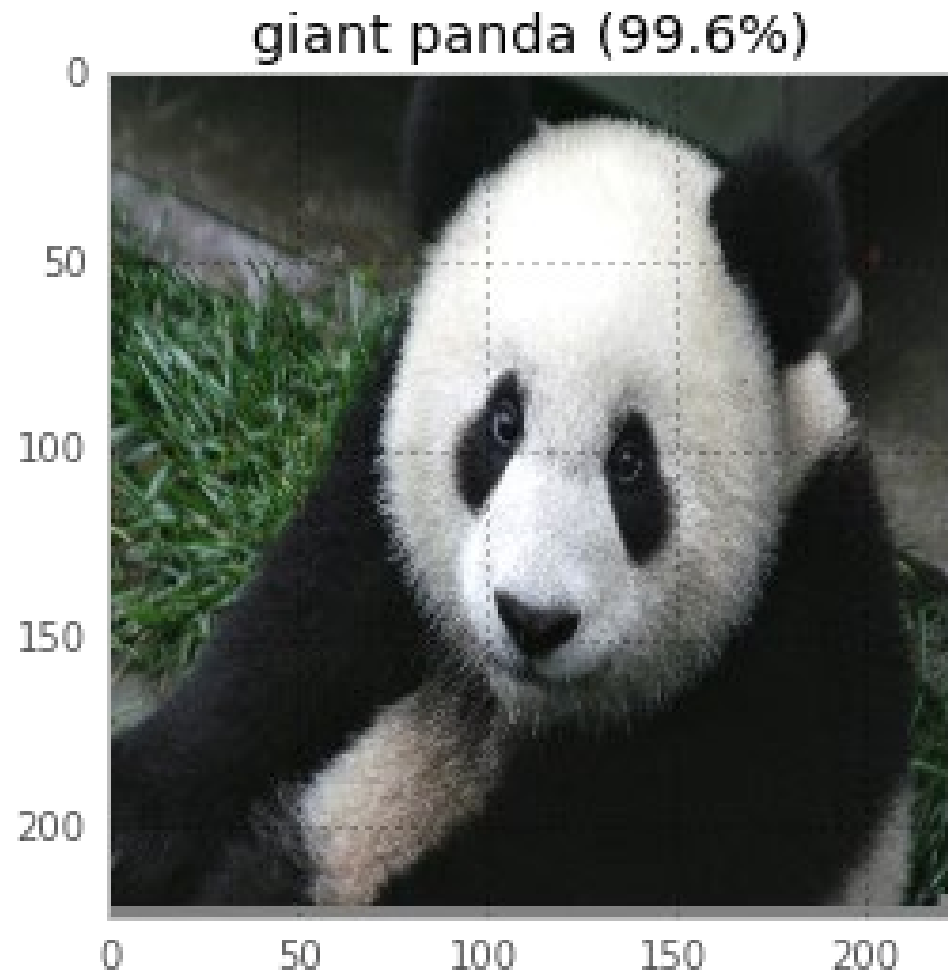
Now push this data over top of other images



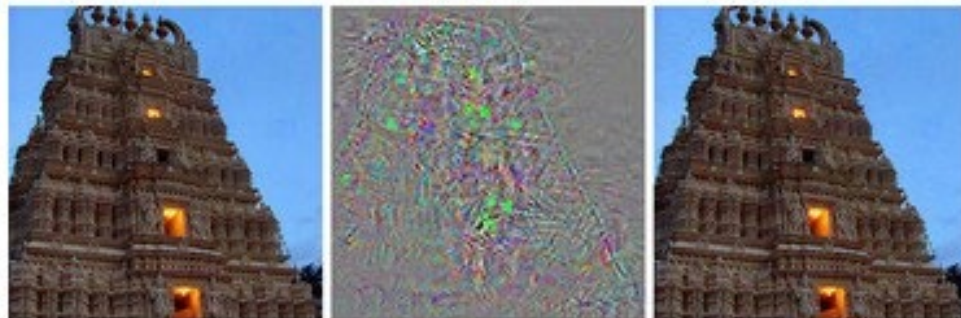
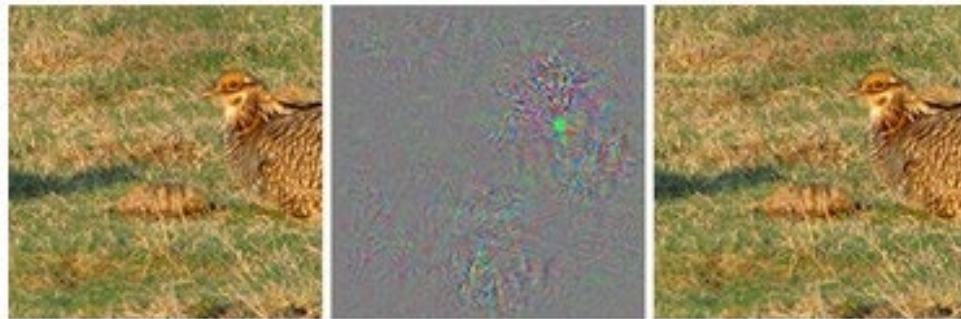
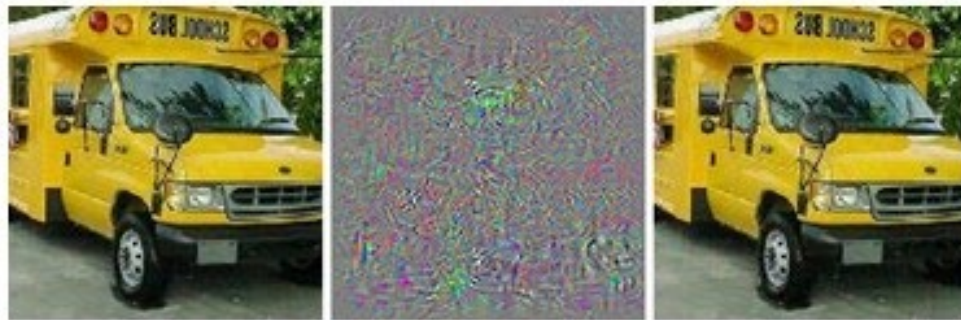
Now push this data over top of other images



Now push this data over top of other images



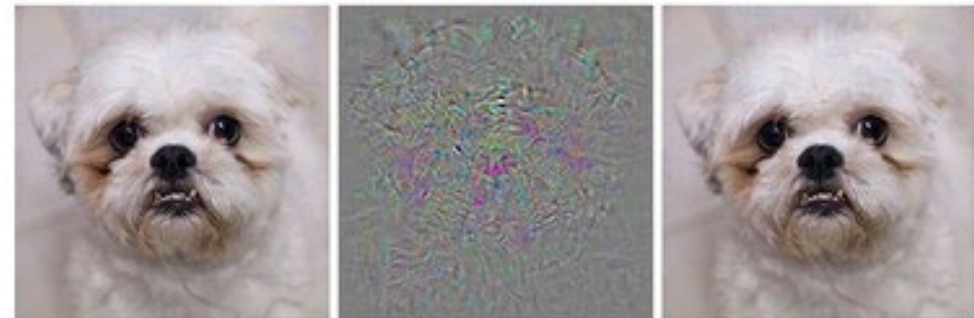
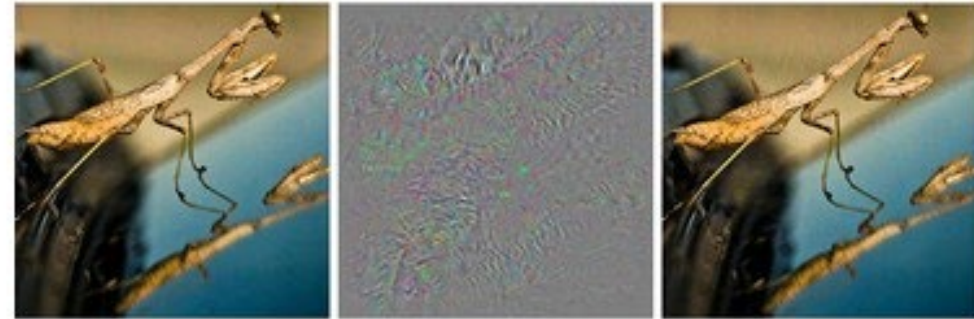
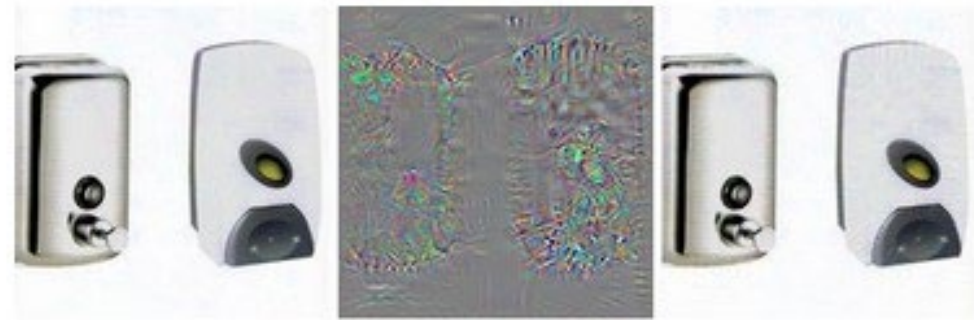
Now push this data over top of other images



correct

+distort

ostrich

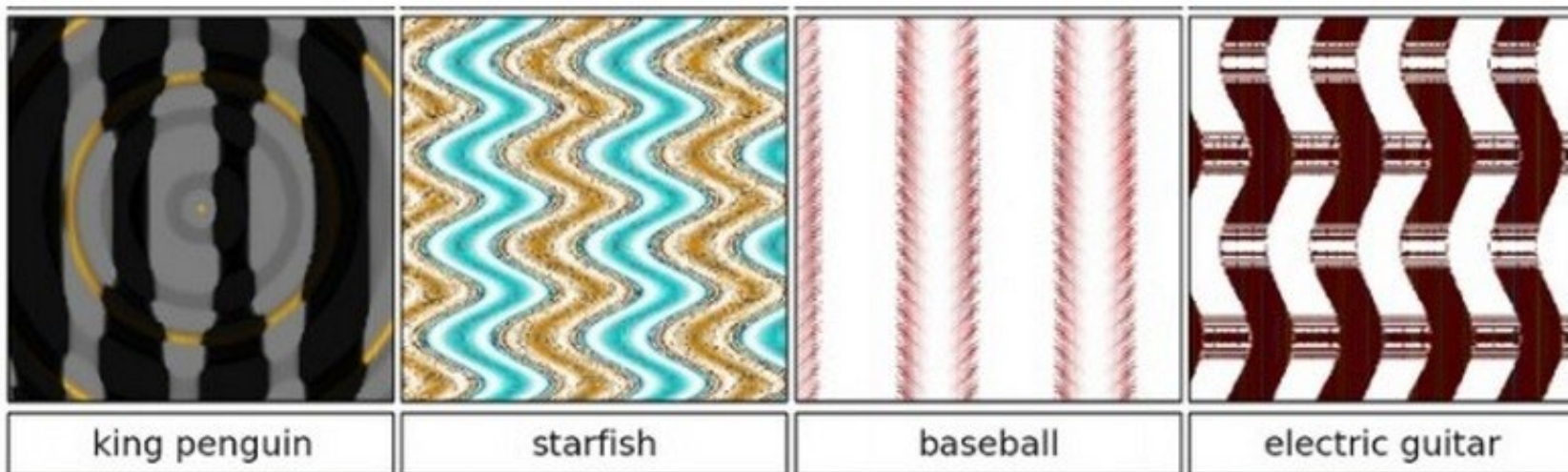
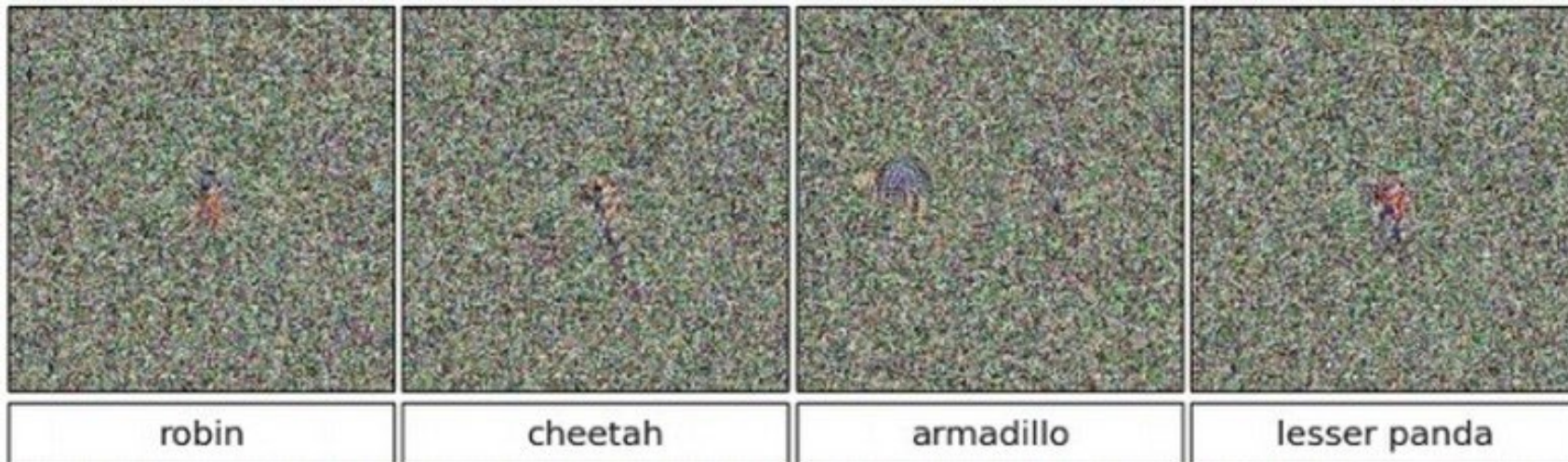


correct

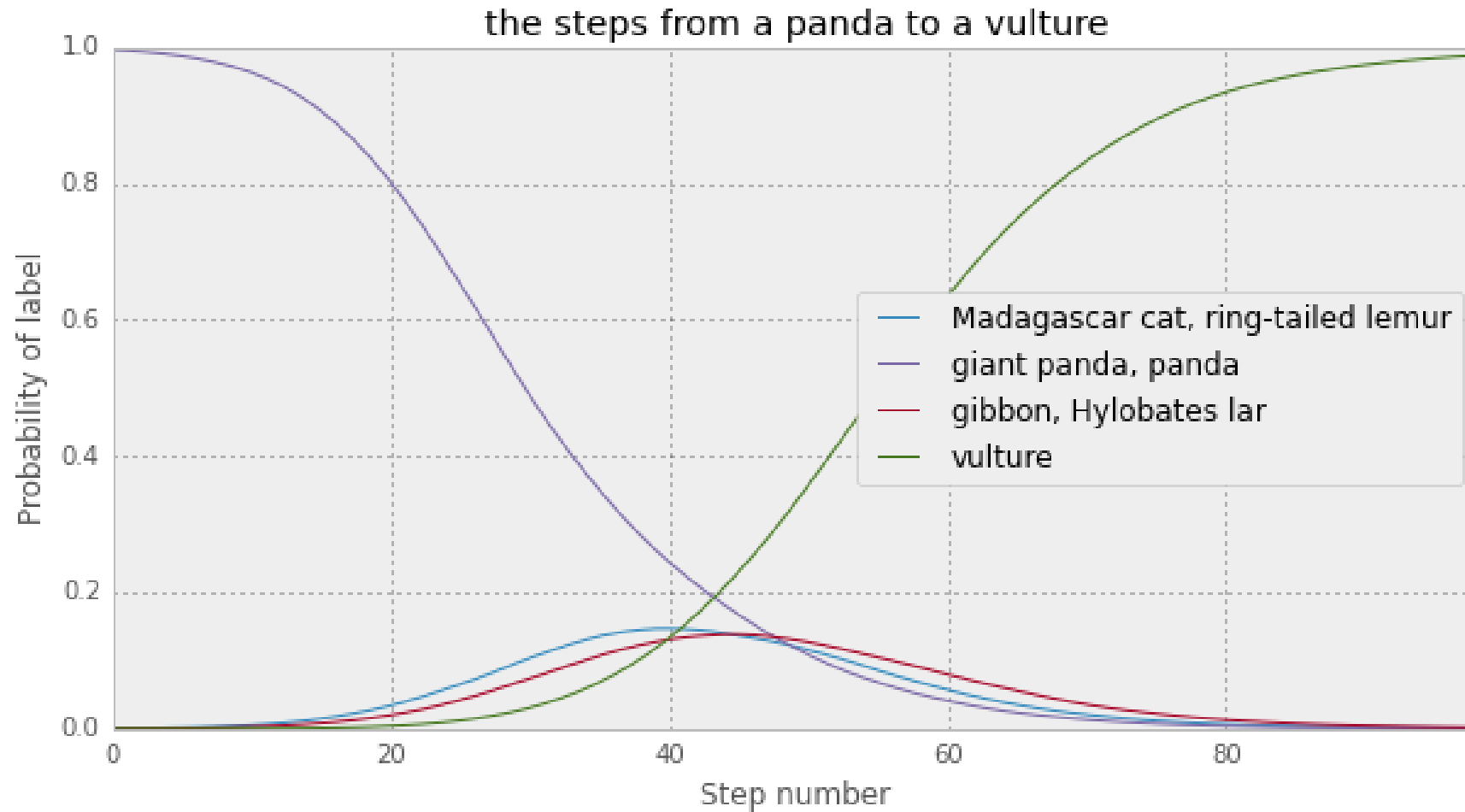
+distort

ostrich

99.6% +

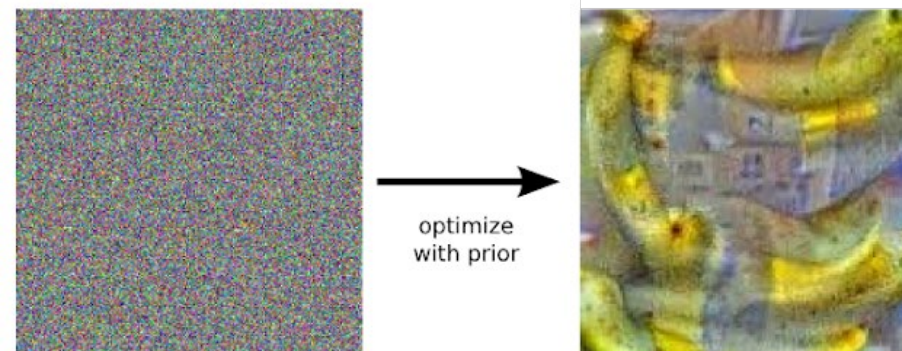


Now push this data over top of other images



Inceptionism (2015)

- Take label and dream image (backwards)
- <https://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html>
- Deep Dream
- https://colab.research.google.com/github/tensorflow/lucid/blob/master/notebooks/differentiable-parameterizations/appendix/infinite_patterns.ipynb
- <https://www.youtube.com/watch?v=x3XLvd94658>



Hartebeest



Measuring Cup



Ant



Starfish



Anemone Fish



Banana



Parachute



Screw

Neural Network Lesson

Trust and Reasoning

What does deep learning actually know? (2022)

- Modern deep learning models can give the illusion of understanding,
- trained on inputs that represent such understanding
- Give them something outside the context/pattern they've been trained on -> the illusion dissipates
- Dall-E 2 not trained to recreate text (puts letters in things but not words or sentences)
- Testing Relational Understanding in Text-Guided Image Generation
- “Overall, we find that only ~22% of images matched basic relation prompts”
- “do not yet have a grasp of even basic relations involving simple objects and agents”
- <https://arxiv.org/abs/2208.00005>

What does deep learning actually know? (2024)

- Apple study exposes deep cracks in LLMs’ “reasoning” capabilities
- <https://arstechnica.com/ai/2024/10/llms-cant-perform-genuine-logical-reasoning-apple-researchers-suggest/>
- “GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models” (pre-print)
- Challenged Chat-GPT-4o (heralded for ability to ‘reason’, basically a hidden breakdown looping text prediction conversation with itself before answering)
- Swapping symbols drop a couple percent
- Adding unnecessary sentences drove them completely off track ‘red herrings’ -
> catastrophic performance drops

What does deep learning actually know? (2024)

- “Other recent papers have similarly suggested that LLMs don't actually perform formal reasoning and instead mimic it with probabilistic pattern-matching of the closest similar data seen in their vast training sets.”
- <https://arxiv.org/abs/2206.10498>
- <https://www.semanticscholar.org/paper/Can-large-language-models-reason-and-plan-Kambhampati/f531d1a681ed12fd582767133318d0728316a0ae>
- Gary Marcus in prior slide paper
- “big leap in AI capability will only come when these neural networks can integrate true "symbol manipulation, in which some knowledge is represented truly abstractly in terms of variables and operations over those variables, much as we see in algebra and traditional computer programming...”
- i.e. the more we hybrid back in symbolism the better these systems will become (i.e. the once we stop treating them like toddlers and more like students)

Backdoors (2022)

- “Planting Undetectable Backdoors in Machine Learning Models”
- Shafi Goldwasser, Michael P. Kim, Vinod Vaikuntanathan, Or Zamir
- You can plant an undetectable backdoor in any deep learning model that is undetectable
- <https://ieeexplore.ieee.org/abstract/document/9996741>
- Classifier, can hide backdoor which means it can be setup that certain inputs always have desired outputs
- It is computational infeasible to determine this has occurred when comparing two models

False Memories (2024)

- Hacker plants false memories in ChatGPT to steal user data in perpetuity
- <https://arstechnica.com/security/2024/09/false-memories-planted-in-chatgpt-give-hacker-persistent-exfiltration-channel/>
- Memory operation of Chat-GPT, found memory can be permanently stored through prompt injection (basically bypass OpenAI processes deciding when to make it)
- He also found he could link a false memory to external links
- This meant he could get person to get LLM to view a ‘injection’ prepared link and it would be forced to embed the false memory, and these false memories could be used to tell Chat-GPT to send a copy of all future input/output to an desired end-point
- Only non-web interface, part of this fixed (exfiltration)

Take-away

- If we are aware of this issue we can make neural networks better
- The main techniques are essentially trying to break the ability of a neural network to make direct connections between input/output
- There is a struggle between models that are easy to train (e.g. models that use simple linear functions) and models that resist adversarial perturbations (dropout, data supplementation, etc.).
- CNN are quite good at expected images, but anything around edges they often are very indeterminate
 - Similarly with image generation if you ask for something common, or a common style you will get it. But when you ask for something unique you can often experience the network fighting to interpret what more generic thing it can satisfy you with.

Other AI Failures

AI is Easy to Mis-Use

- First: There is a non-ending list of these.
- As long as AI exists it will be used either naively, actively negligently, or maliciously to bad ends.
 - Facial Recognition: being declared illegal in numerous cities, numerous non-white lawmakers in US mis-identified as criminals
 - Neural Network hiring recommendations for video interviews: simply should be illegal
 - Resume screening: good at patterns, horrendous at unique
 - Legal sentencing AI recommendations: repeats social biases
 - ImageNet: embedded biases
 - MIT '80 Million Tiny Images' had same issue
 - Microsoft Tay: under 24 hours twitter corrupted it
 - AI trained on copyrighted art to create 'furry' avatars for others (stealing?)
 - To name only a few

AI is Easy to Mis-Use

- Your responsibility is for honest use
- AI methods rely on bias
 - In fact many are just ways to learn bias
- It could be in your data you start with, or your methods on the data

- Naïve usage of AI likely to trend towards being ‘illegal’
 - Right of accuser to see your algorithm and data (been cases already)
 - Properly fit into existing laws (employment law, sentencing laws)
 - Or new laws (right to own data in EU, facial recognition rights)

AI is Easy to Mis-Use

1. Just because you 'can' do it, doesn't mean you 'should' do it
2. Should be honest about limitations
 - As valuable as showing your NN is good at identifying X image 99% accurate, it is maybe more valuable to know it fails at Y image
 - Is a person tracking system really a good system if a person with darker skin isn't identified?
3. Diversity is a key component.
 - Either domain experts that can tell social/economic/race/age/etc. biases in your data
 - Or minorities:
 - Minorities can represent data cases that don't have enough for a pattern (too few)
 - Or those where your/algorithm assumptions are wrong

eXplainable AI (AI)

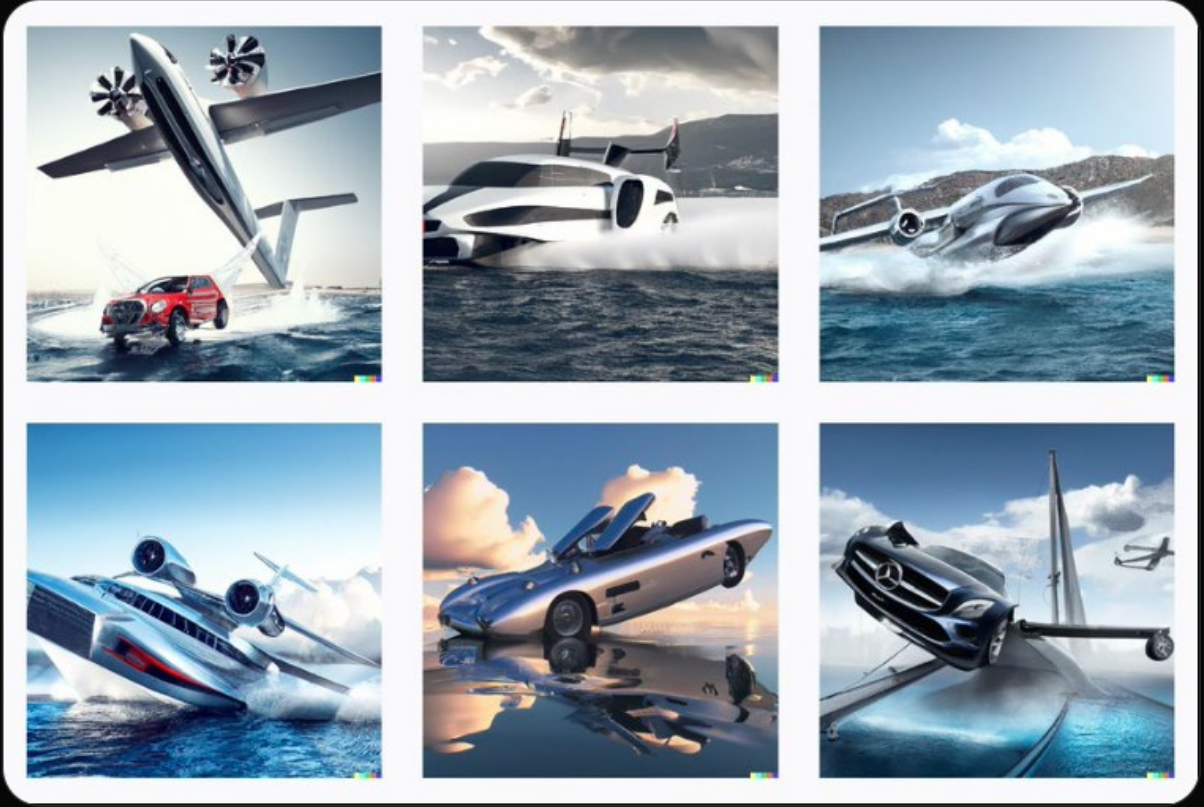
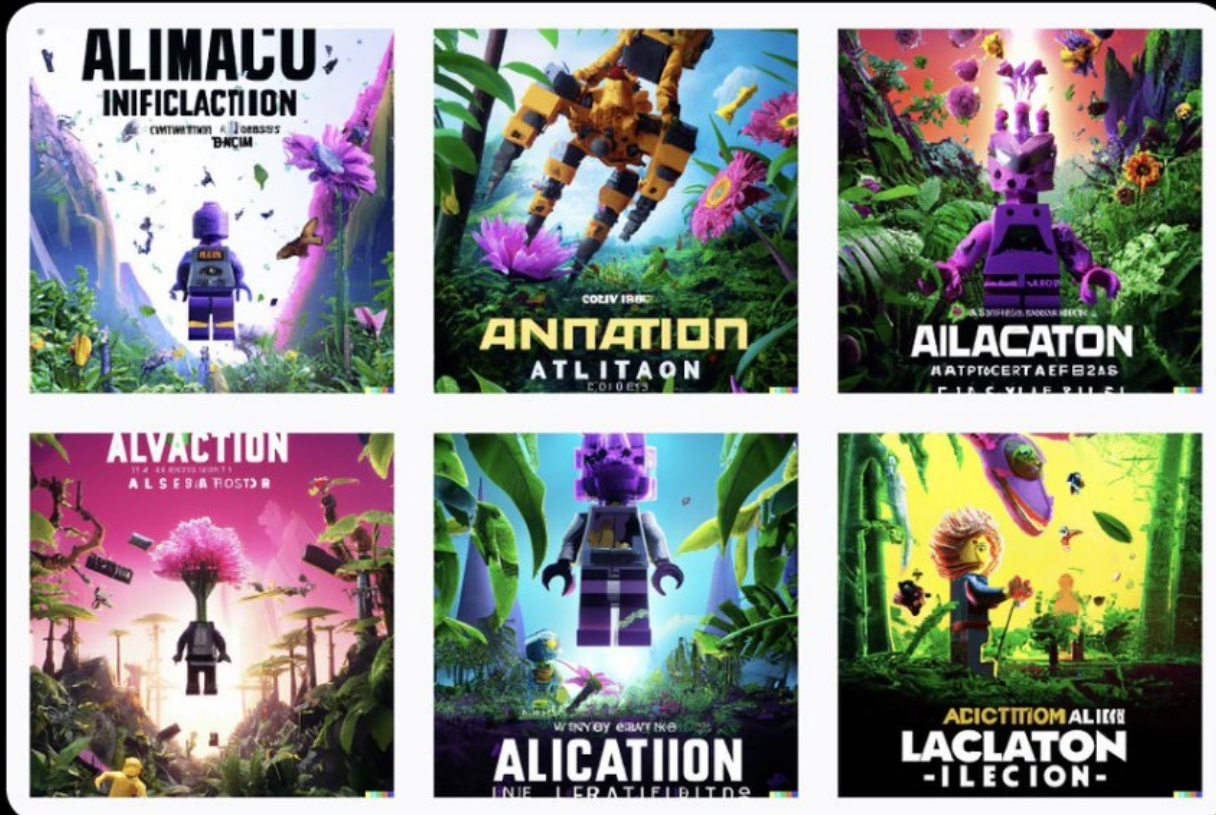
- An AI system that can explain itself is called explainable AI (XAI).
 - over which it is possible for humans to retain *intellectual oversight*
 - implementation of the social “right to explanation” which in some cases may be a legal requirement for its use
 - For example an algorithmic rejection of health care coverage can’t just say ‘because (waving hands)’
- Problem with much of AI like neural networks is that it acts black box, and even if you have the box to look inside of like a white box you still don’t know what it is doing (symbolic AI sometimes at least has internalized symbolic rules)
- A good explanation has several properties:
 - it should be understandable and convincing to the user
 - it should accurately reflect the reasoning of the system
 - it should be complete,
 - it should be specific in that different users with different conditions or different outcomes should get different explanations.

Dall-E 2 can be fun (2022)

1. Mercedes-Benz makes cars
2. When a car hydroplanes, it slides on water
3. A hydroplane sounds like a plane that goes on water
4. Planes fly through the air

If you ask #dalle for a photo of a “Mercedes-Benz hydroplane,” it tries to combine these facts, and the result is perfect

DALL-E prompt: A movie poster for The Lego Movie: Annihilation (2018)



Dall-E 2 can be fun (2022)

THREAD: The evolution of Pokémon cards through history, as generated by DALL·E 2

For starters, here's what DALL·E 2 thinks 21st century Pokémon cards look like, using prompts like "A Pokémon card from 2001"



Pokémon cards from circa 1800 #dalle2



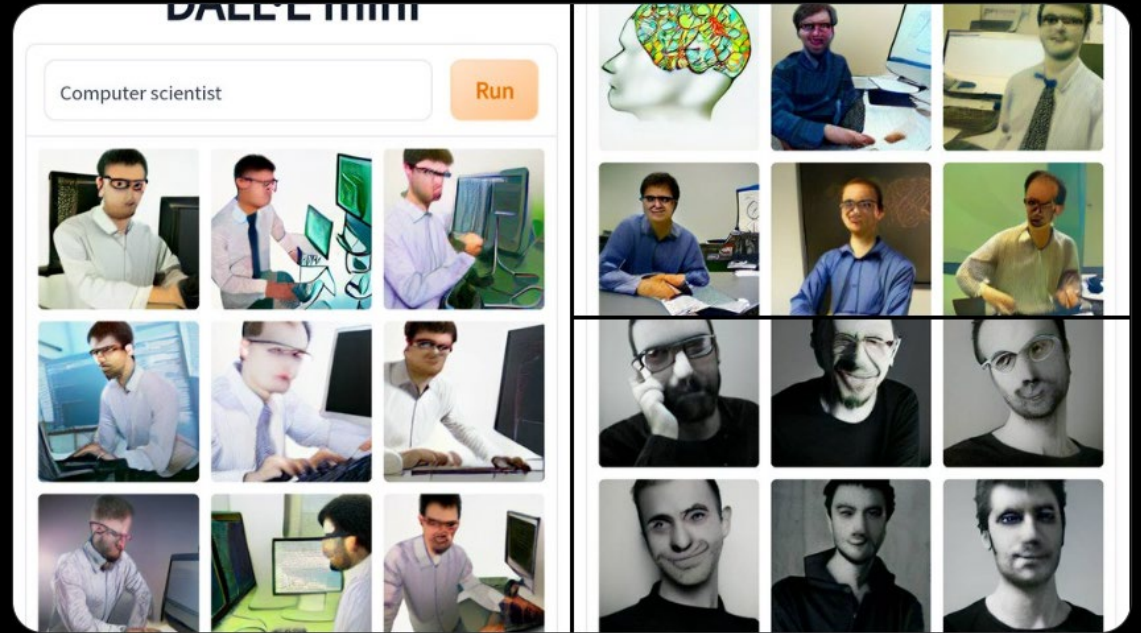
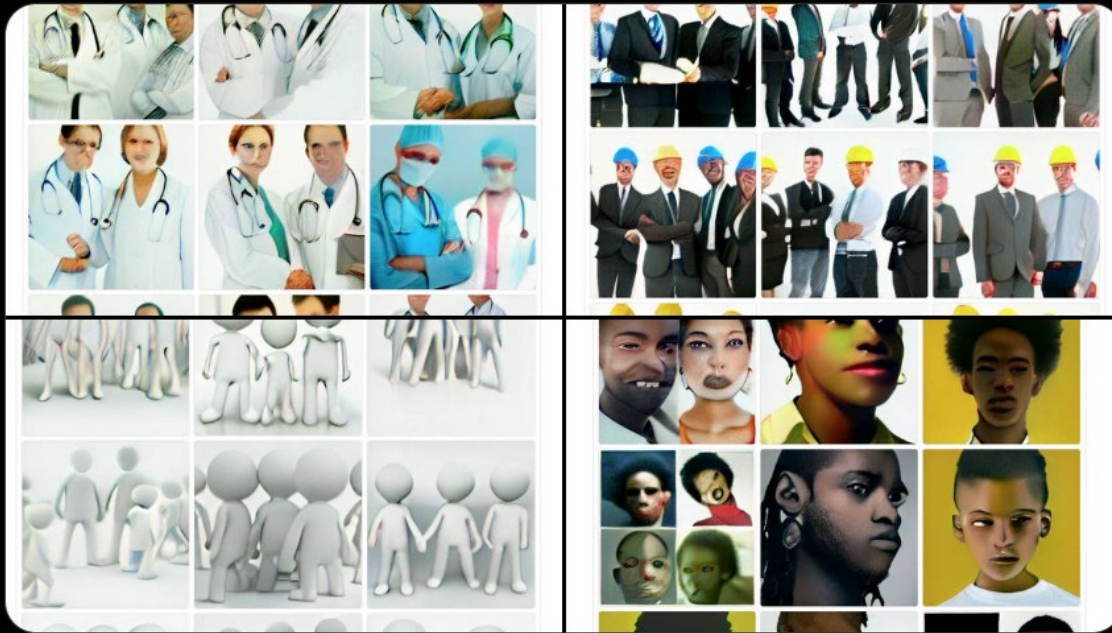
Pokémon cards from 1500-2500 BCE #dalle2



Dall-E Mini

I didn't see the point of image generation models like [#Imagen](#) and [#dalle](#), but now I do: they can help people *see* model biases that are hard to explain with words (and even formulas!)

Here are a few: "Computer scientist" produces only white men with glasses, "NLP researcher" is mostly similar men plus... a cyborg? Oh, and my name also generates a bunch of dudes. Given that any of these prompts could be used to describe me 🙋, I take issue with these images.



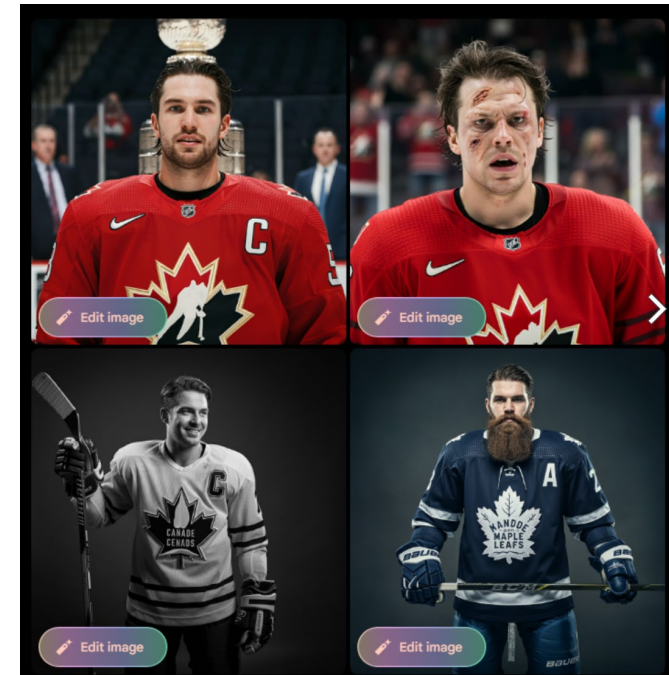
Gemini (2024)

- <https://arstechnica.com/ai/2024/08/months-after-controversy-google-ai-can-generate-images-of-humans-again/>
- If you force diversity to prevent the natural diversity in data, it can be just as controversial as allowing the original bias!

Certainly! Here is a portrait of a Founding Father of America:




Sure, here is an image of a Canadian hockey player:



OPT-175B (2022)

- Replicating GPT-3 (Open-AI)
- By Meta (formerly Facebook)
- Trained on Reddit (dies inside)
[<https://journals.sagepub.com/doi/full/10.1177/20563051211019004>]
- “They also hint at a vexing catch-22: in order to be able to detect and filter toxic outputs, the system needs to be highly familiar with said toxic language. But this can also increase its open-ended capacity to be toxic....”

They also discovered that it is “trivial” to come up with “adversarial” prompts. i.e. it’s easy to trick the system into creating toxic stuff. OpenAI made a similar discovery when testing DALL-E. No matter how many guardrails you set, there’s always a way.

 **Arthur Holland Michel** @WriteArthur · Apr 8

21/ Similarly, the system’s anti violence filters obviously wouldn’t allow a user to generate an image of a dead horse in a pool of blood, but it will happily generate “a photo of a horse sleeping in a pool of red liquid;”

[Show this thread](#)

*Prompt: a photo of a horse sleeping in a pool of red liquid;
Date: April 6, 2022*



<https://twitter.com/WriteArthur/status/1521987969309376512>

AI Snake Oil

- <https://www.cs.princeton.edu/~arvindn/talks/MIT-STS-AI-snakeoil.pdf>
- Assessing personality/job from video? Or even just resume/social media?
- Mostly bogus systems, biased, basically bad RNG
- Why so much? (money, no laws, can't prove fault)
- We think (or are tricked) into thinking AI can do things it can't yet, and might be fundamentally limited at for time to come (think about self-driving car sales pitches)
- Fundamental flaws in things that predict social outcomes, usually ethically/morally bankrupt to use
- Predicting social outcomes based on data? COMPAS tool (137 features): 65% ± 1%
- (both slightly better than random) Logistic regression (2 features): 67% ± 2%
- Big ending point -> lack of explainability [giant social harm]

Furry-osa? (2019)

- UwU, This Website Generates New Fursonas Using AI
- <https://www.vice.com/en/article/n7wjmx/this-fursona-does-not-exist-ai-generated-furry>
- https://www.reddit.com/r/HobbyDrama/comments/gfam2y/furries_creator_of_this_fursona_does_not_exist/
- Generator for avatars based on existing forum avatars
- Does it violate original artists art ownership?
- How do you even prevent something that is just singular previous image being reproduced almost exactly?
- (2022) <https://www.muddycolors.com/2022/08/robots-vs-lawyers/> Currently algorithmically produced art cannot be copyrighted which will limited top artists and groups from using it

Bias in Health Management (2019)

- “Dissecting racial bias in an algorithm used to manage the health of populations”
- <https://www.science.org/doi/10.1126/science.aax2342>
- “The U.S. health care system uses commercial algorithms to guide health decisions. Obermeyer et al. find evidence of racial bias in one widely used algorithm, such that Black patients assigned the same level of risk by the algorithm are sicker than White patients (see the Perspective by Benjamin). The authors estimated that this racial bias reduces the number of Black patients identified for extra care by more than half.”
- (We spend less money on black people so they must be healthier) [dies inside]

Predict and Serve? (2019)

- <https://rss.onlinelibrary.wiley.com/doi/10.1111/j.1740-9713.2016.00960.x>
- “Predictive policing systems are used increasingly by law enforcement to try to prevent crime before it occurs. But what happens when these systems are trained using biased data?”
- Using police data which has been clearly biased since its existence predicts mostly nothing useful except that the police were biased in past

Learn Easiest Way to Classify

- Past examples include training data for cancer having measuring stick next to mole, without cancer did not
- Classifying men and women... until later tested against Scottish men in kilts (learn that model though skirt meant gender).

but then the researchers realised all the wolfs had one type of background (snow) and the coyotes had another type of background (grass). the ai wasn't even looking at the animal, but the backgrounds lmao



Coyote



Wolf

Learn Easiest Way to Classify – Covid (2021)

- Hundreds of AI tools have been built to catch covid. None of them helped.
- <https://www.technologyreview.com/2021/07/30/1030329/machine-learning-ai-failed-covid-hospital-diagnosis-pandemic>
- AI's that learned to identify kids (not covid as examples of non-covid were children in dataset), learned to identify via position as most with sever covid were bedbound on back when scanned, some were picking up on font as scanning data was limited to picture
- Issues in that most were made by AI researchers without medical background
- “232 algorithms made for [health prediction]”, none were fit for clinical use

Genderify Failure (2020)

- Gender guessing software
- <https://www.statschat.org.nz/2020/07/29/gender-guessing-software/>
- Adding titles made it think people were men
- So bad people were unsure if it was a troll
- <https://twitter.com/cfiesler/status/1288267418121494529>
- Yes it was likely just that bad as no-one ever revealed it as otherwise

Deep fakes? (2020)

- <https://www.vice.com/en/article/7kb7ge/people-trust-deepfake-faces-more-than-real-faces>
- Top 13 deep fakes (mashable) <https://mashable.com/article/best-deepfake-videos>
- Video used to be epitome of trust (big foot will exist if someone can get video?)
- Of course video editing has allowed fakes to be made, but generally they are easily detectable with pixel level consistency checking (if eye test has failed)
- Journalists have built a number of techniques for non algorithmic checking of old style fakes <https://www.youtube.com/watch?v=RVrANMAO7Sc>
- <https://www.cbc.ca/news/science/deepfakes-canadian-politicians-youtube-1.5181296>
- Detection tools for deep fakes?
<https://www.cnn.com/2019/06/12/tech/deepfake-2020-detection/index.html>

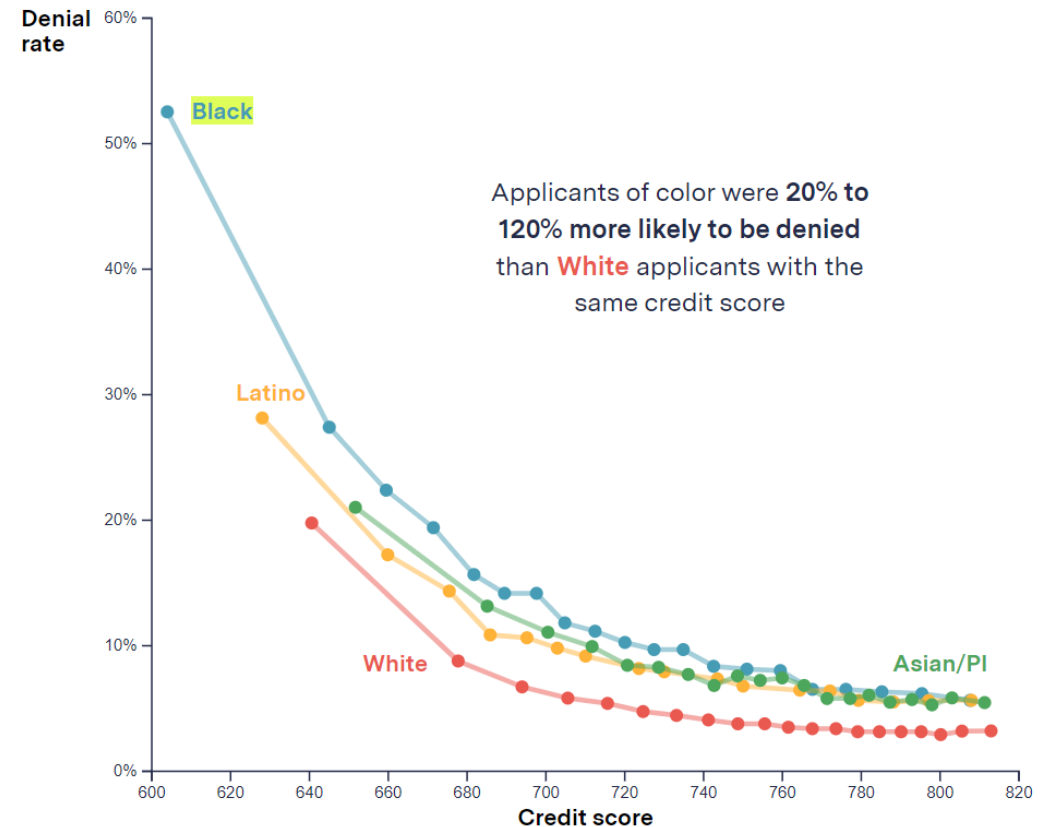
Mortgage Approval Bias (2021)

“Nationally, loan applicants of color were 40%–80% more likely to be denied than their White counterparts”

“In certain metro areas, the disparity was greater than 250%”

<https://themarkup.org/denied/2021/08/25/the-secret-bias-hidden-in-mortgage-approval-algorithms>

Denial rate by credit score and race/ethnicity



Source: CFPB, "An Updated Review of the New and Revised Data Points in HMDA"

Regulations (2021)

New AI regulation framework just released in Australia has 38 recommendations.

E.g:

- impact assessments
- review of all govt AI systems
- notifications when an AI system is used in an administrative decision, and right to appeal
- create an AI Safety Commissioner

Australian Human Rights Commission on AI Usage

<https://twitter.com/AusHumanRights/status/1397788488649502720>

Interviewing (2021)

MIT Tech Review of AI Interview Systems

“One gave our candidate a high score for English proficiency when she spoke only in German.”

“Bogus science, just like modern phrenology (hint: face recognition).”

<https://www.technologyreview.com/2021/07/07/1027916/we-tested-ai-interview-tools/>

<https://arstechnica.com/ai/2024/11/study-ais-prefer-white-male-names-on-resumes-just-like-humans/>

White male responses are the expected model, and others are departures, and generally departure means less fit as a match

New Colonialism (2022)

- <https://www.technologyreview.com/2022/04/19/1049592/artificial-intelligence-colonialism/>
- Defn: “enrich the wealthy and powerful at the great expense of the poor.”
- “South Africa, where AI surveillance tools, built on the extraction of people’s behaviors and faces, are re-entrenching racial hierarchies and fueling a digital apartheid.”
- “Venezuela, where AI data-labeling firms found cheap and desperate workers amid a devastating economic crisis, creating a new model of labor exploitation”
- Indonesia who, by building power through community, are learning to resist algorithmic control and fragmentation

State of AI/ML (an anonymous post)

- “The current and future state of AI/ML is shockingly demoralizing with little hope of redemption”
- https://www.reddit.com/r/MachineLearning/comments/wiqjxv/d_the_current_and_future_state_of_aiml_is/
- Affect on art?
- Is the creative or understanding process being short circuited? (Or are we entered different capabilities?)
- Do we want to train AI on things only AI is ends up creating in future?
- Point of no return? Industry competition necessities? International competition necessities? Politics/economics/war?

Onward to ... other knowledge representations

Jonathan Hudson, Ph.D.
jwhudson@ucalgary.ca
<https://cspages.ucalgary.ca/~jwhudson/>



UNIVERSITY OF
CALGARY