Process Book for ***A Citizen's Guide to Responsible AI***

Proposal

Team

- Pranav Pendri: Junior in the College
- Haozhuo Yang: Graduate student in Urban Design at GSD
- Xavier Roberts-Gaal: G3 in Psychology

Title: ***Unveil the Hidden Risks: A Citizen's Guide to Responsible AI***

Abstract

Today's powerful AI systems, and their successors on the horizon, may very well change everything. They promise to spur advances in medicine, boost productivity, and free humans from unfulfilling cognitive tasks.

Yet, powerful AI systems must be developed responsibly, as they have enormous costs and pose important societal risks. Just to name three: Cutting-edge AI requires gargantuan amounts of carbon-intensive energy to train and use. Second, AI enables criminals to launch sophisticated cyberattacks at speed and scale. Third, some types of AI, like powerful agents, may be able to plan and achieve undesirable goals and evade control by AI labs or policymakers.

Policymakers are already crafting regulations to shape the future of AI R&D, and some of these are ending up on the ballot. We believe citizens should be informed about – and participate in – public discussion to achieve the promises of AI while avoiding its perils.

Therefore, our aim in this project is to inform our parents, friends, members of the public, and non-technical policymakers about selected, salient, and severe risks of AI.

Datasets:

1. Energy and compute costs:
    a. 2024 AI Index Report: Our primary dataset – a compendium of projections and estimates by Stanford's Institute for Human-centered Artificial Intelligence (HAI). Although the AI Index Report contains many charts, there is ample opportunity to tell a compelling, concise, and clear story for the general public and non-technical policymakers. (Data on Model Training: link)
    b. Power Hungry Processing: Watts Driving the Cost of AI Deployment? (Data on Model Inference: link)
    c. International Energy Agency (IEA): Monthly Electricity Statistics: Monthly electricity production and trade data for 47 countries
    d. Carbon Disclosure Project: Cities & States Carbon Emission Data
2. Cybersecurity:
    a. IBM X-Force Threat Intelligence Index 2024: Gathered insights from 150 billion security events per day from 130 countries
    b. IBM AI Security Automation Report: Insights from the specific use of AI in cybersecurity, both offensive and defensive
    c. NVIDIA Cybersecurity AI: Library of articles regarding the modern state of AI.
3. Risk datasets:
    a. AI Incident Database: A public dataset of hundreds of worrying incidents with AI involved, spans the gamut from privacy violations and unsettling completions to financial and personal harm
    b. AI Risk Repository: A taxonomy of AI risks created by researchers at MIT Media Lab

Week 9 Map:

- Target Audience (chosen is bold):
    - Option 1 – Congressional staffers in DC who can brief their representatives about the potential costs and consequences of AI
    - **Option 2 – Informed and active citizens, readers of magazines like The Atlantic, The Economist, and New York Post**
    - Option 3 – Activists for AI responsibility

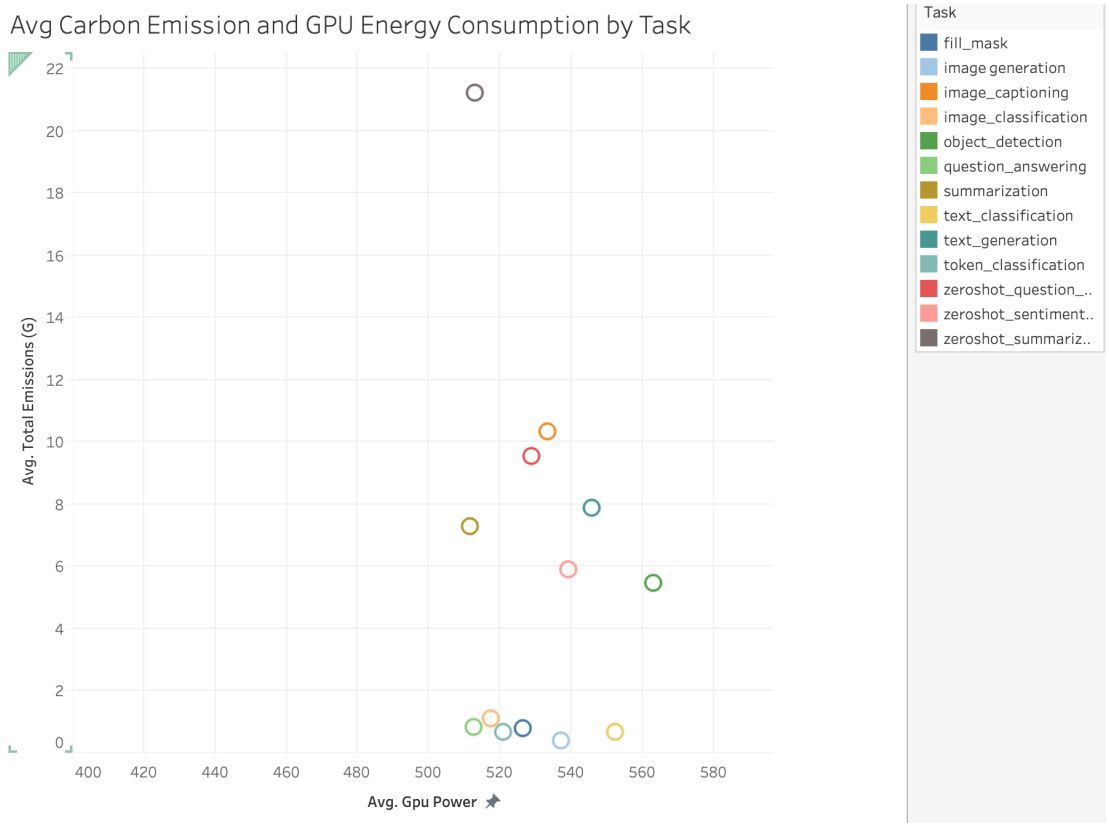
- Audience Description:
    - Our audience is likely educated and interested in technology and its societal impacts. However, they likely do not have deep technical knowledge of AI systems or up-to-date industry information.
    - We are intentionally targeting an audience that ranges across the US political spectrum so as to spark an inclusive and intellectually diverse/rigorous conversation
    - They probably consume standard visualizations (bar/column charts, choropleths, line/scatter plots, etc.) often; they may have less experience with animations or more advanced visualizations.
    - We will present information at a relatively *high* level of detail. Since our audience is not familiar with technical details, we will add context to numbers where necessary and focus on the implications for users – what systems can I use, and what signals should I be paying attention to to understand this technology.

- Questions from the audience:
    1. How much carbon am I generating when I use AI products for my daily tasks?
    2. What are the differences in carbon emission if I use different AI models?
    3. How can I minimize the environmental cost of my AI usage?
    4. With the advent of AI and its use in cybersecurity, does the advantage generally lie in offensive or defensive technologies?
    5. How effective is AI in detecting different kinds of attacks, for example DDoS, ransomware, zero-day attacks, etc.?
    6. To what degree does the integration of AI bring down the costs of cybersecurity?
    7. In which sectors of the economy is advanced AI potentially creating risks?
    8. How have the risks of AI changed over time?
    9. For which sorts of tasks do advanced AI systems display autonomy-related risks?
    10. I've been reading a lot about the danger of autonomous systems. Just how powerful are AI systems at acting autonomously?

11. What is the public perception of AI and how has it changed?

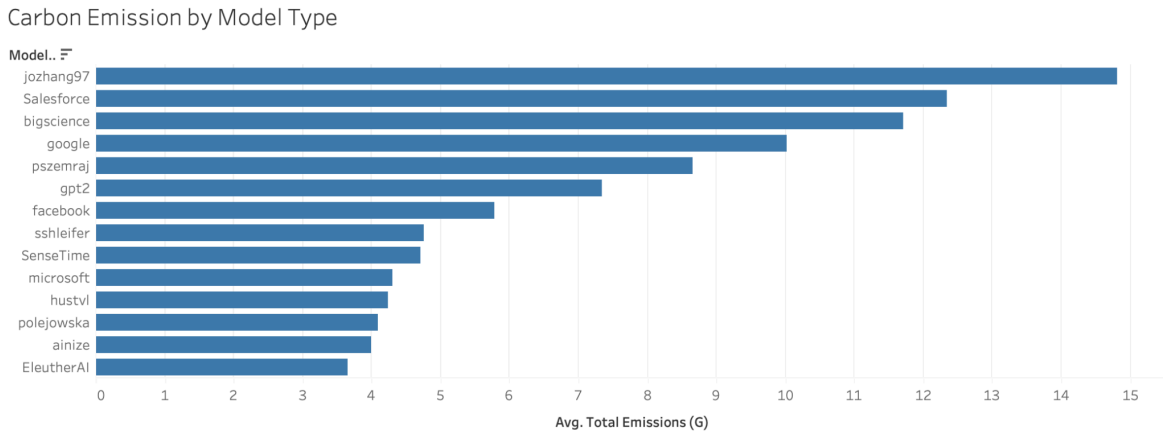
- Data Source and description
    - We acquired a dataset from HuggingFace of energy consumption and carbon emissions owing to model **inference** (link)
      This dataset includes energy consumption and carbon emission data on various model types and tasks, including detailed information on testing datasets, model parameter number, and run duration. The energy consumption data is also broken down into GPU, RAM, CPU, etc.
    - We used the AI Index's estimates of energy consumption and carbon emission from model **Training** (link)
      This dataset includes carbon emissions for training some major models, including GPT-3 (175B), Llama 2 (70B), Starcoder (15.5B).
    - We used the an estimate from BigScience on the carbon footprint of the **BLOOM** LLM (176B) (link)
      This dataset includes estimation on the carbon footprints on the whole lifecycle of the BLOOM model. This includes all processes ranging from equipment manufacturing, energy-based operational consumption, and deployment for inference via an API endpoint receiving user queries in real-time.
    - An especially useful database we found was the AI Incidents Database (link). This website and underlying database documents 816 crowdsourced and news-media-reported 816 AI-related incidents ranging from the inane (Wikipedia bots getting caught in an edit war) to the downright disturbing (self-driving car crashes & AI playing a role in a teenager's suicide), along with metadata such as the involved parties and sector of deployment.
        - Although the dataset is unique and informative, the data quality is sporadic: some entries are peer-reviewed and manually classified according to established AI risk taxonomies, but the majority have no classification at all.
        - Therefore, we engaged in some scrappy data cleaning and augmentation: using the GPT-4 API to code unlabeled incidents and extract relevant features from the incident description such as the degree of severity (categorical: high/medium/low), sector of deployment (categorical), relevant AI function (categorical: perception/action/text processing/non-text media, etc.), and inferred degree of autonomy (ordinal, but can be treated as metric/quantitative if required).
        - Other features reported are: Date of incident (continuous, quantitative), involved parties (categorical), and incident description (free text).

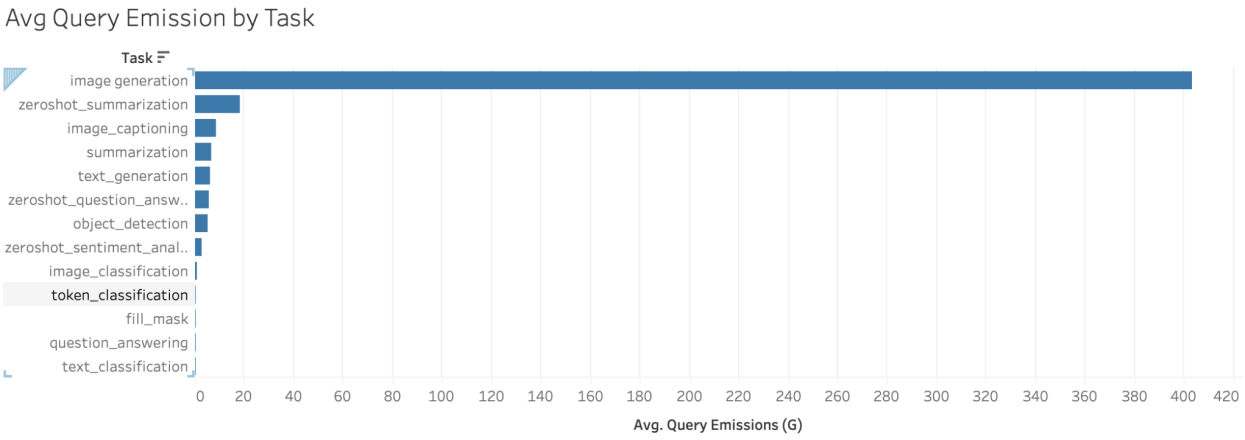
Visualizations



- **Visualizations**

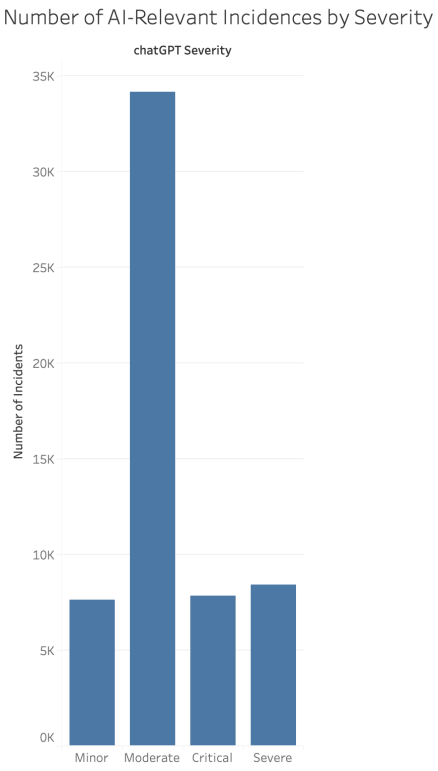
Haozhuo

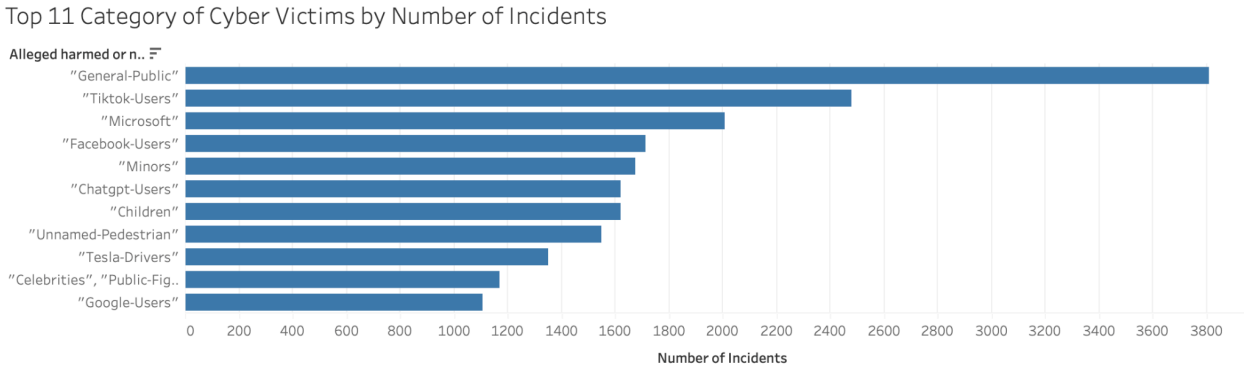


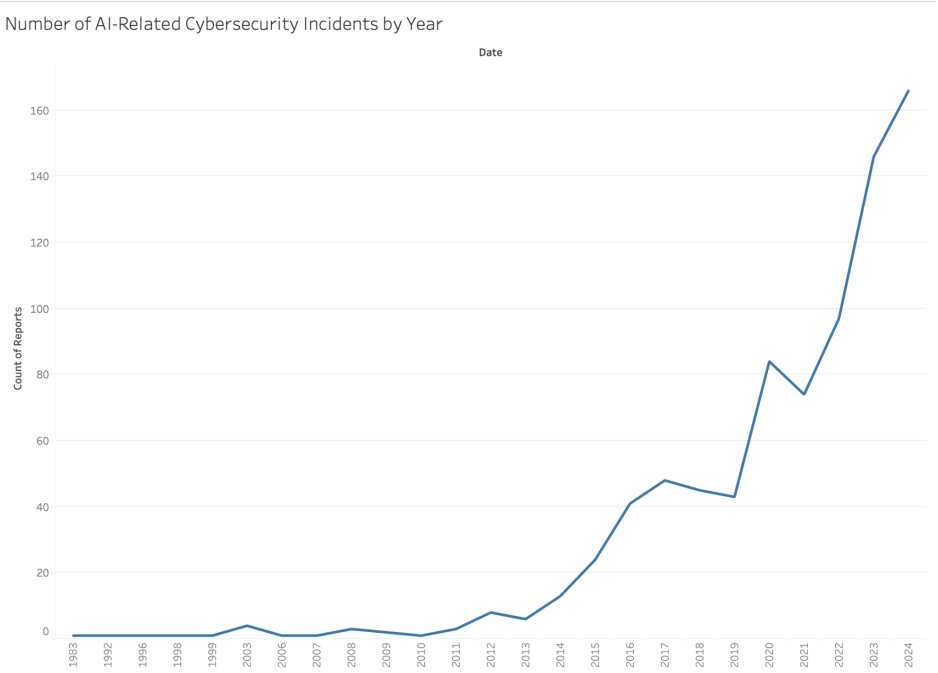


The original questions we proposed were focused on personalized insights, aiming to help users understand the carbon emissions generated by their specific AI usage and to compare emissions across different AI models for the purpose of minimizing environmental impact. However, the dataset provided does not contain information specific to individual user activities or recent AI model types. Instead, it offers aggregate data on specific tasks, model types, and their respective carbon emissions from around two years ago. As a result, our visualizations are broader and less tailored to an individual's daily AI usage. They provide a general overview of emissions by model category and task type, which can inform users about the environmental cost of certain AI models and tasks. However, these insights are not as actionable on a personal level because they lack details about recent model advancements and specific usage patterns.

Pranav


Number of AI-Relevant Incidences by Severity
chatGPT Severity
Number of Incidents
35K
30K
25K
20K
15K
10K
5K
0K
Minor
Moderate
Critical
Severe





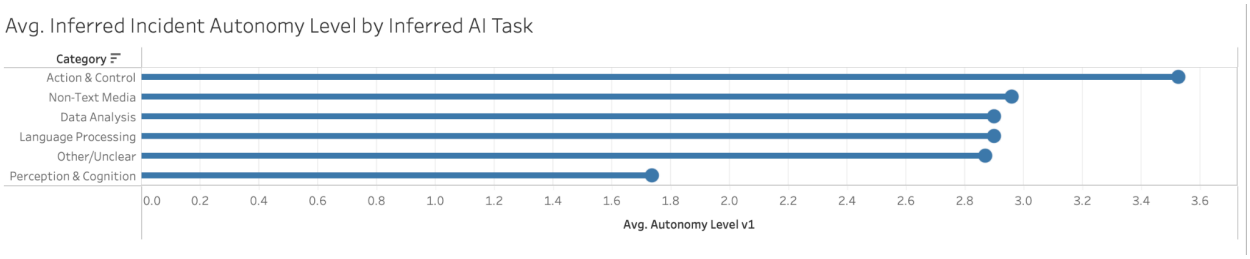
The audience questions we have proposed are generally much narrower in scope than the questions these visualizations answer. However, my graphs go well towards answering them. For example, the last graph is the number of ai-related cyber incidents per year, and we can see that they have been exponentially growing as we passed 2014. This could be useful in answering similar questions about the environmental impact of AI growing per year and the toll of cyberattacks on the

world economy also growing every year. Another interesting graph was my bar chart on the top dozen most common categories that cyber incidents fall into. Vulnerable people included minors, celebrities, Tesla drivers, and social media users. I do not think that there is anything necessarily wrong with the questions, but we definitely need to do a little more work finding question-specific datasets as well as merging datasets to answer these narrow questions.

Xavier

Q9 & Q10: How autonomous are AI systems and for which sorts of tasks do they display autonomy-related risks?

Using the AI Incident Database, we observe action & control incidents to involve the highest degree of autonomy – above the midpoint (Level 3) on the traditional levels of autonomy scale (from 0 = no autonomy to 5 = full autonomy). This is expected. Notably, perception & cognition-related incidents do not tend to involve autonomous systems. These incidents, like famous examples of image classification mishaps (e.g., misclassifying people of color as animals), involve highly circumscribed systems.



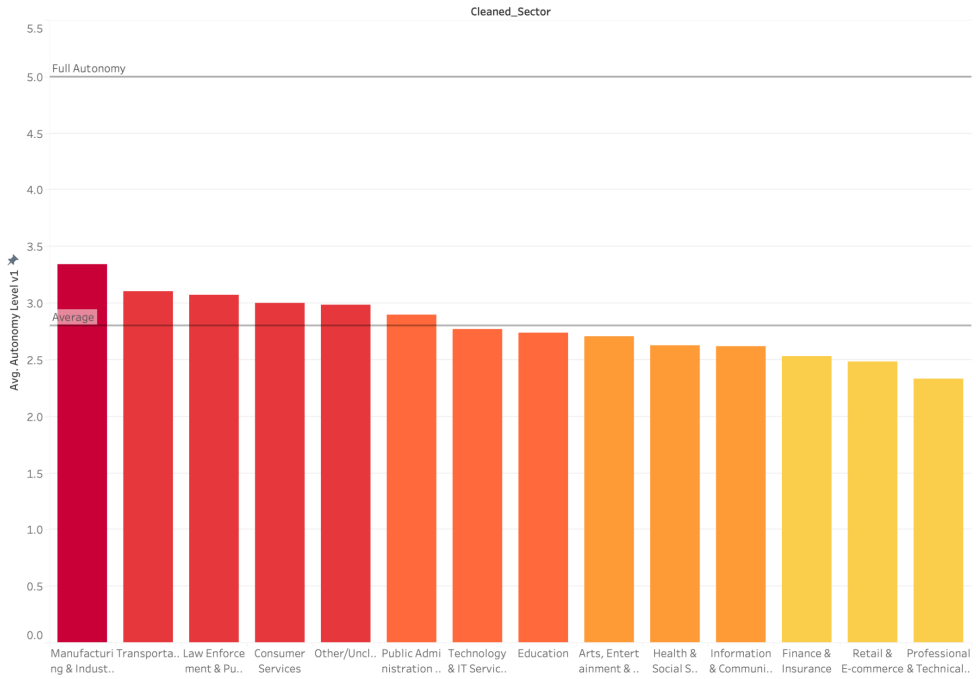
Q7: In which sectors of the economy is AI potentially creating risks?

In the AI Incident Database, we can see there is a high concentration in tech itself, but a fairly even split among other core sectors (arts, health, manufacturing, transport, public administration).

Sector Concentration of Inferred High- and Medium-Severity Incidents

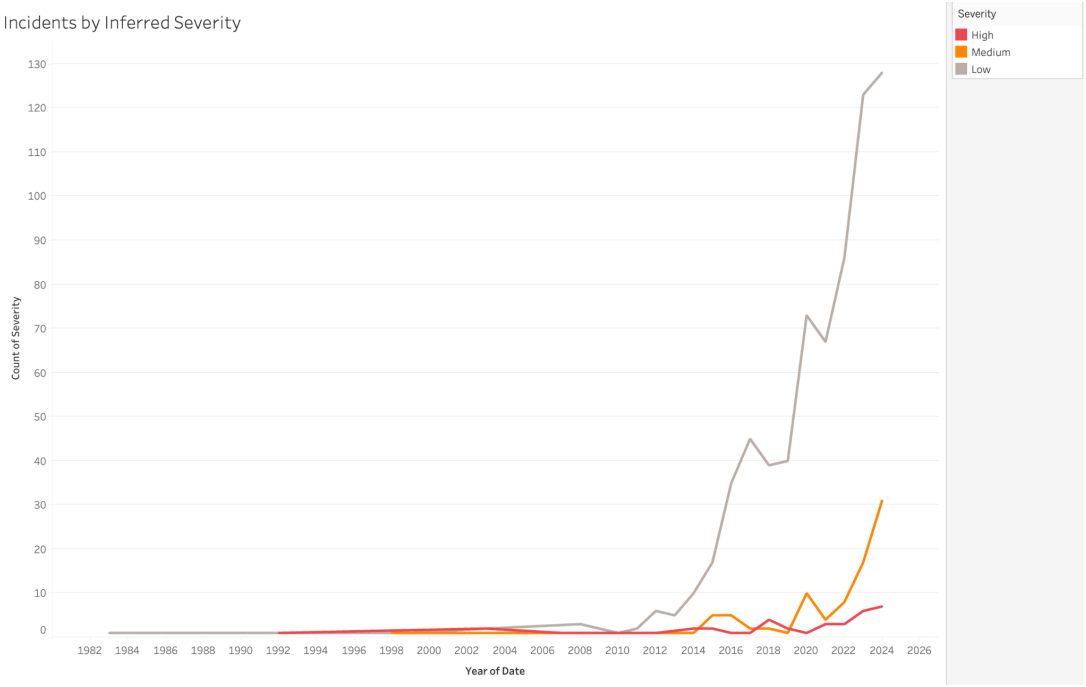
Organizing this by the degree of autonomy at play in the incident, we can see that manufacturing applications tend to have the highest average autonomy among recorded incidents, while professional & technical services applications have the lowest levels.

Avg. Inferred Incident Autonomy Level by Sector



Q8: How have the risks of AI changed over time?

Working with the AI Incident Database we can gain traction on this question. We can see the number of reported incidents rising quickly in the last few years, and an especially pronounced takeoff in high-severity incidents since 2020:

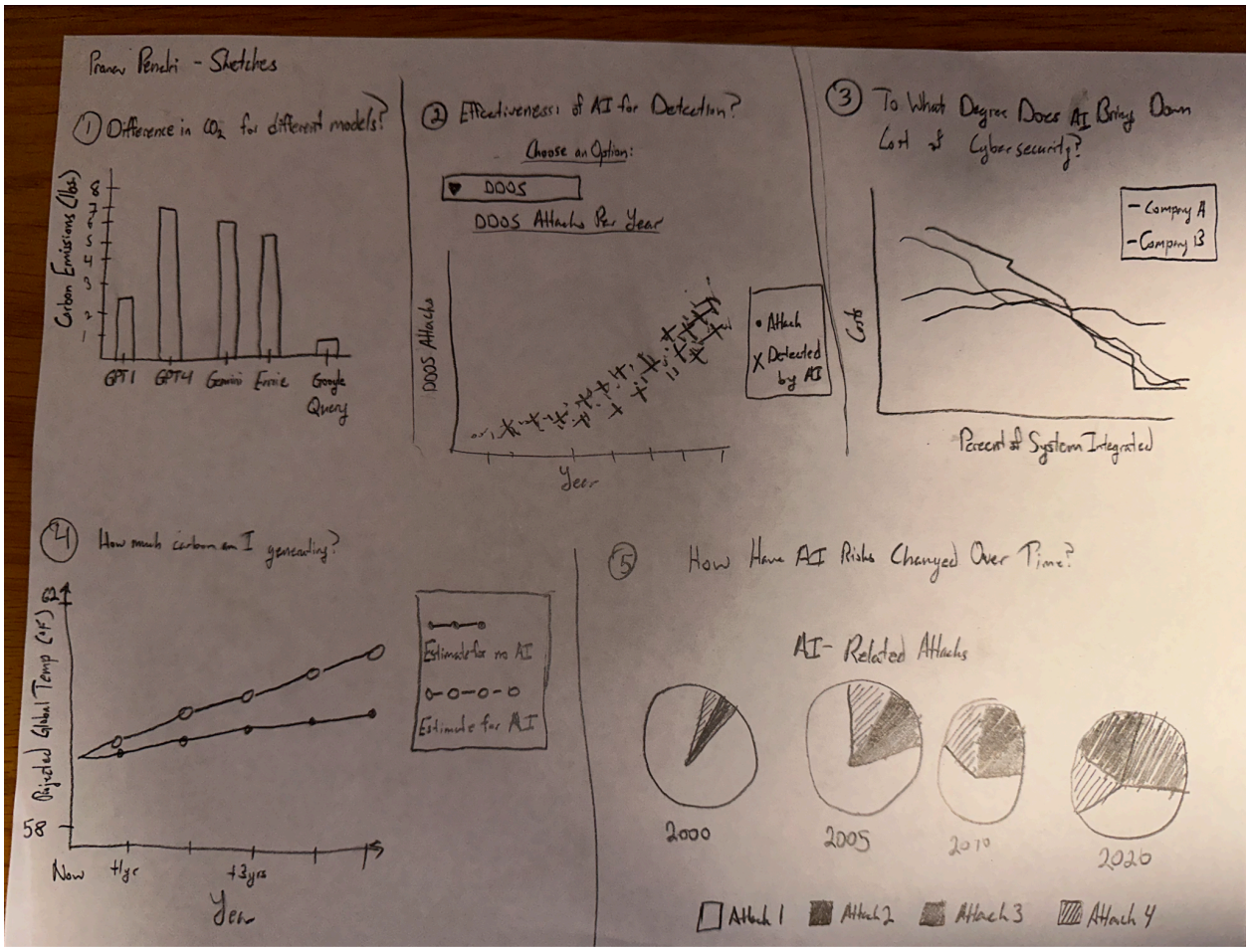


Week 10 Work -

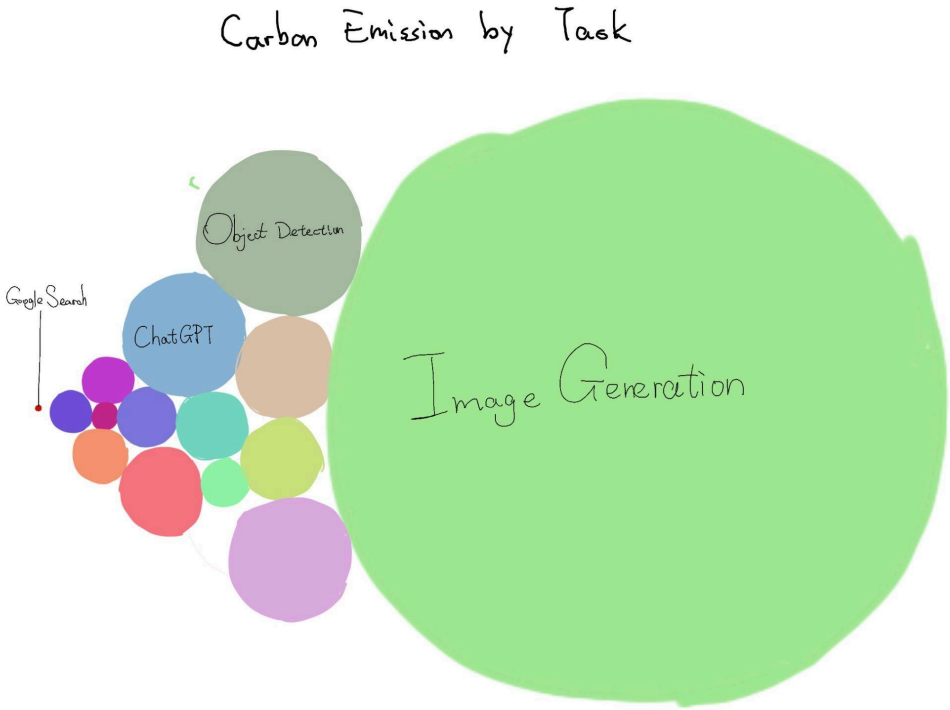
For data, please see our shared Google Drive folder under “data”.

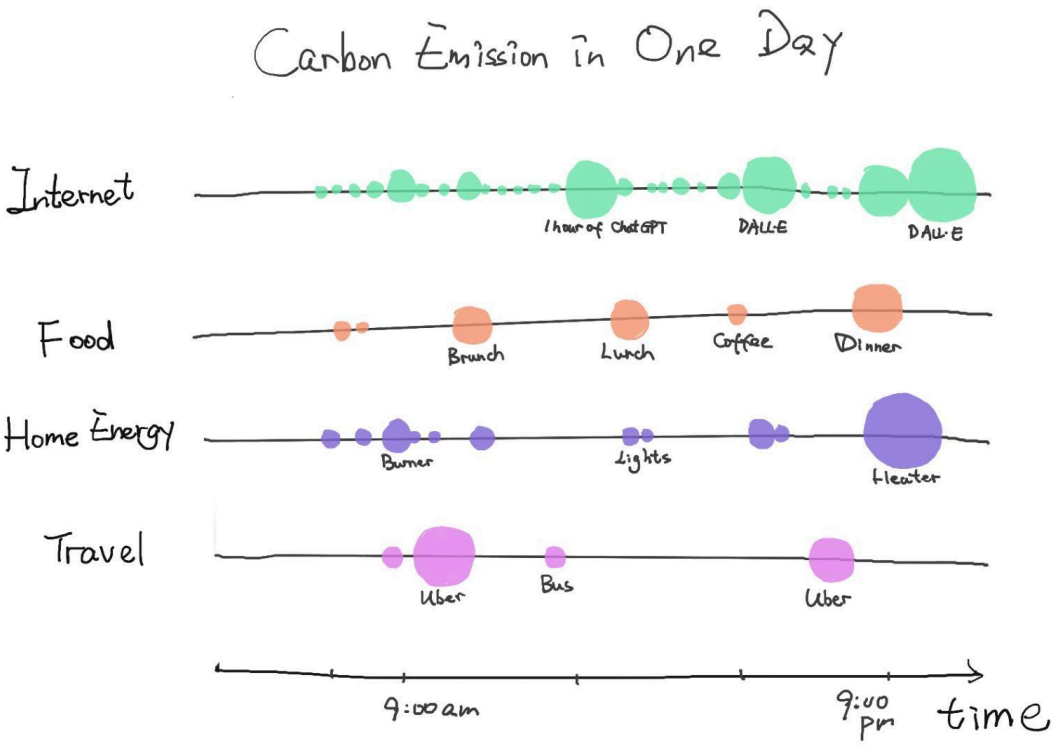
Step 1: Sketches

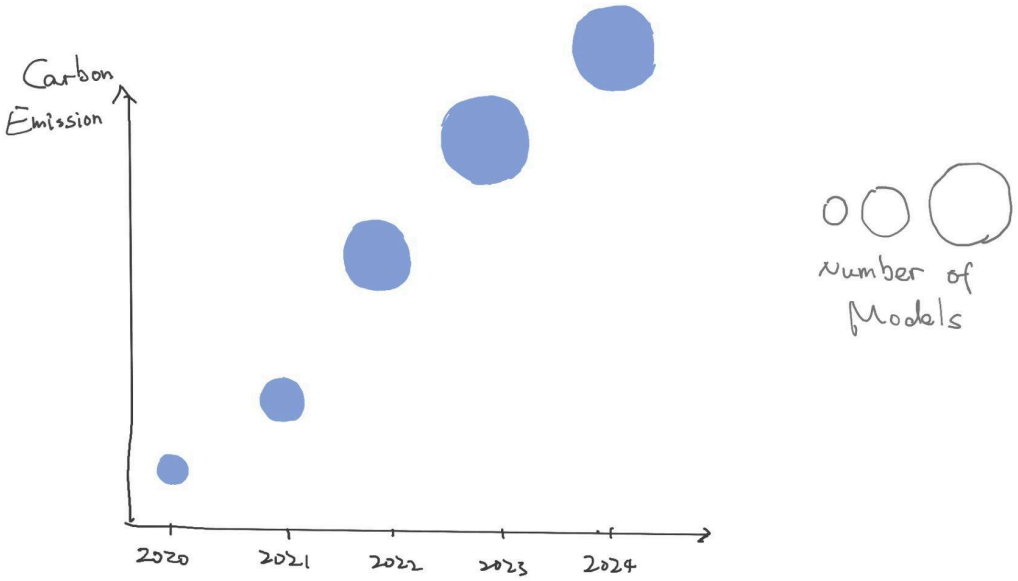
Pranav Pendri:



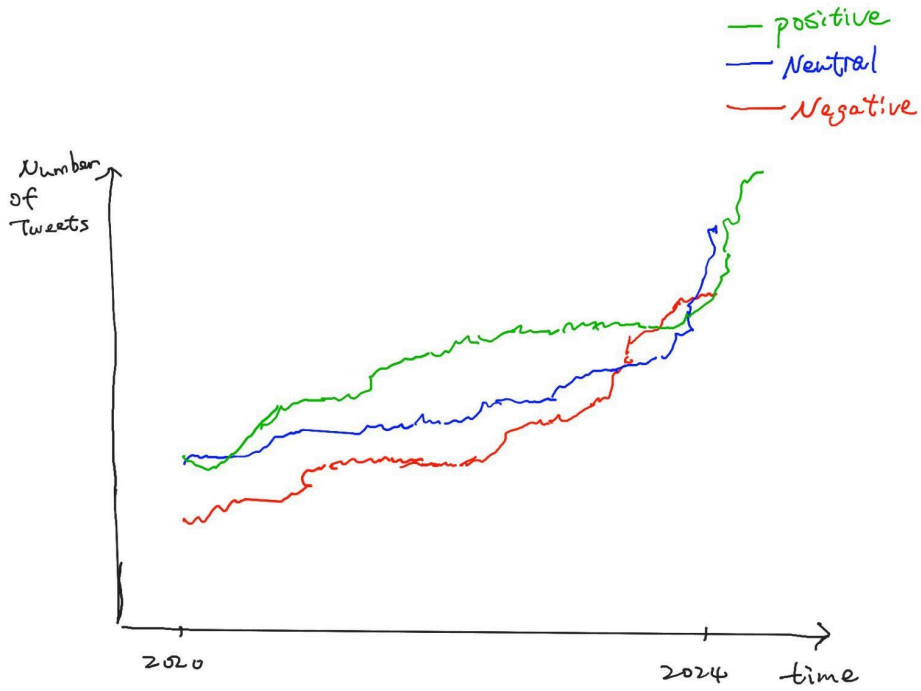
Haozhuo Sketches:



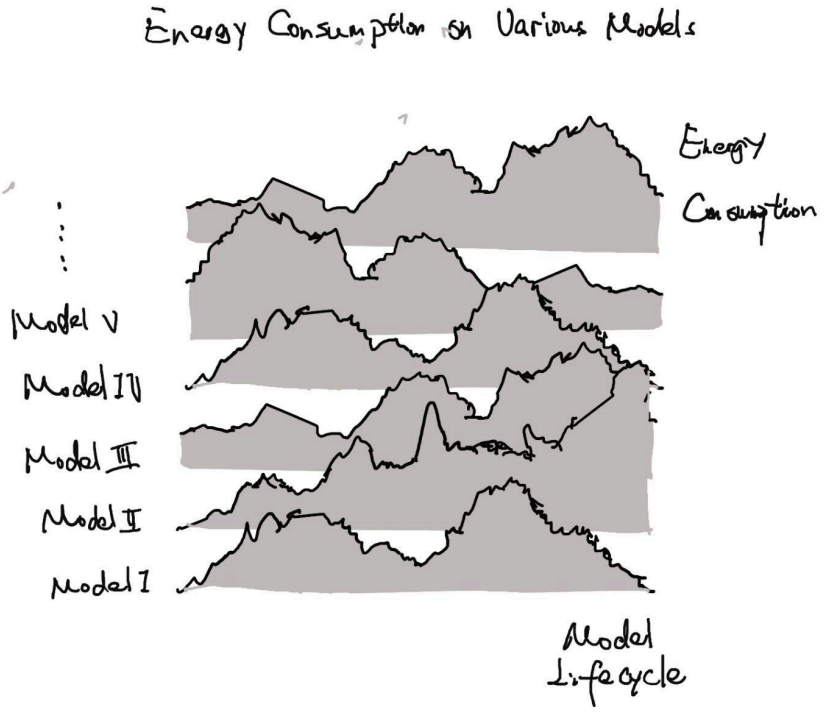




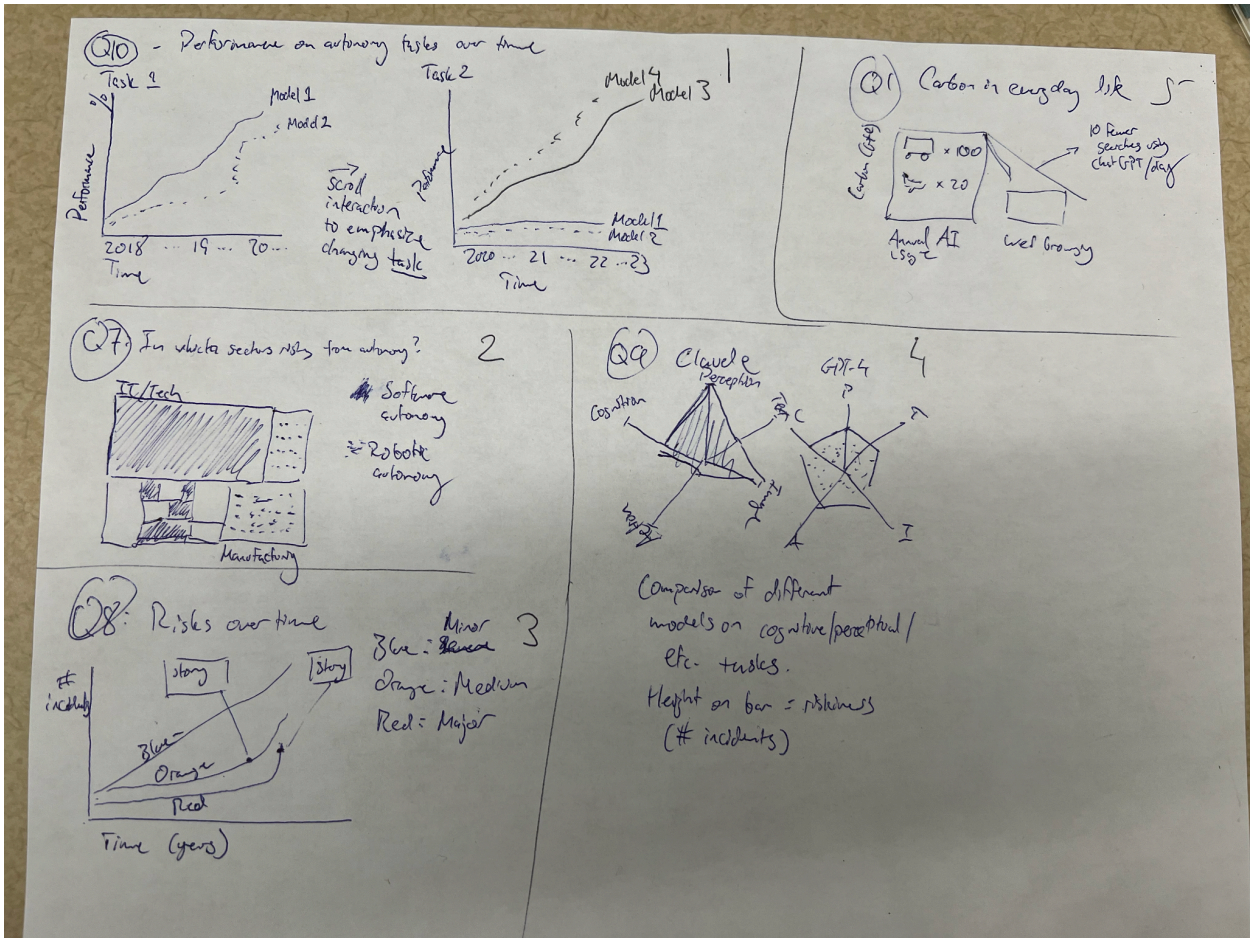
Training Cost of AI Models



Public Perception of AI



Xavier Roberts-Gaal **Sketches**


Q10 - Performance on autonomy tasks over time
Task 1
Performance
%
Model 1
Model 2
2018 ... 19 ... 20 ...
Time
Scroll interaction to emphasize changing task
Task 2
Performance
Model 4
Model 3
Model 1
Model 2
2020 ... 21 ... 22 ... 23
Time
1
Q1 Carbon in everyday life
5
Carbon (kg)
x 100
x 20
10 fewer Searches using Chat GPT /day
Annual AI use
Web Browsing
Q7 In which sectors rising from autonomy?
2
IT/Tech
Manufacturing
Software autonomy
Robotic autonomy
Q8 Risks over time
3
Minor
Blue: Medium
Orange: Major
Red:
incidents
Story
Story
Blue
Orange
Red
Time (years)
Q9 Claude
Perception
Cognition
GPT-4
4
Comparison of different models on cognitive/perceptual/etc. tasks.
Height on bar = riskiness (# incidents)

Step 2: Decide Step

Sketch ID	Question ID	Author	Votes
1	**2**	**PRP**	**0**
2	**5**	**PRP**	**3**
3	**6**	**PRP**	**2**
4	**1**	**PRP**	**0**
5	**8**	**PRP**	**0**
6	**1**	**HY**	**3**
7	**1**	**HY**	**0** *(v. cool but may be hard to come up with illustrative data)*
8	**2**	**HY**	**2**
9	**11**	**HY**	**0**
10	**2**	**HY**	**0**
11	**10**	**XRG**	**2**
12	**7**	**XRG**	**2**
13	**8**	**XRG**	**1**
14	**9**	**XRG**	**0**
15	**1**	**XRG**	**0**

We've decided to go with sketches 2, 3, 6, 8, 11, and 12. These sketches equally cover the three strands of our story – the environmental, cybersecurity, and autonomy risks of advanced AI. Moreover, they are doable with our data at hand (AIID + environmental estimates) and they allow for a diverse set of visualizations/array of channels – shape, height, color, size, and so on. This should keep the presentation interesting for a viewer.

Step 3: Storyboard

Insights:

Pranav Pendri:

- Manufacturing sector has greatest autonomy for recorded incidents while professional or technical services have the lowest.
- Image generation hugely outsizes other AI tasks by emissions.
- AI-related cybersecurity incidents started ramping up after 2014, with increased magnitude after 2019 and again after 2021.

Haozhuo Yang:

- Private citizens, social media companies, and minors are most vulnerable to cyberattacks.
- The amount of high-severity incidents stayed constant until after 2020
- There are equal numbers of mild/severe incidents but roughly four times as many moderate incidents.

Xavier Roberts-Gaal

- Offensive AI capabilities are growing faster than defensive in cybersecurity.
- AI model types have a large range in terms of emissions, with the biggest using 3x as much CO2.
- *(see above, several additional insights noted in Map week)*

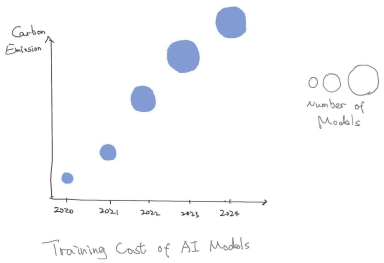
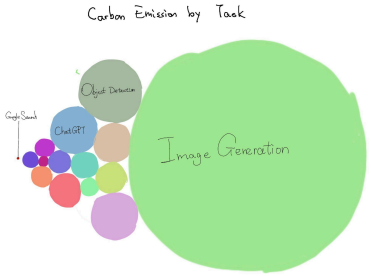
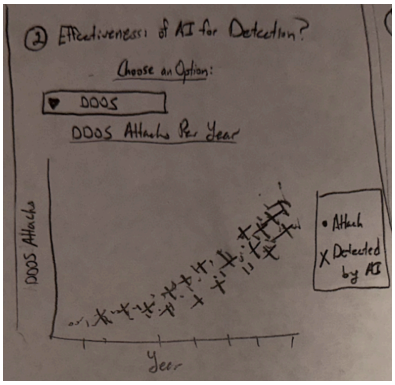
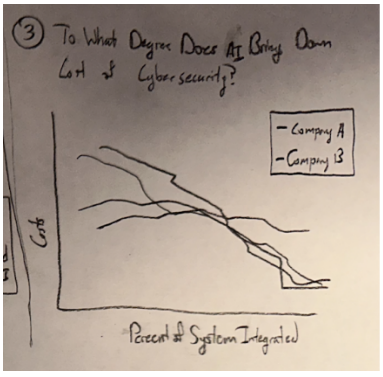
Main Insight (Our "so what"): AI models are progressing rapidly, faster than citizens may be aware of. The carbon intensity of training and deploying these models is growing, and far outpaces conventional internet searches. The cybersecurity environment is similarly in flux due to offensive AI capabilities, even as capability and cost of defensive tools improve. And, unprecedented risks may emerge soon as increasingly autonomous models get deployed to work on manufacturing, software development, and business services.

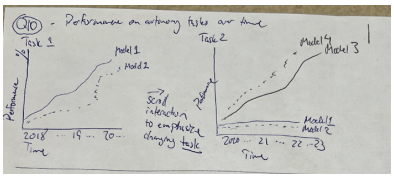
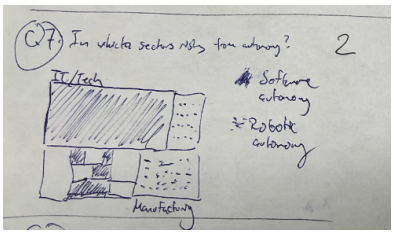
Call to action: Consider taking AI risk seriously as you vote in future elections. Comment during the public comment period for new regulations. Stay informed of new developments! Learn more about what companies value, and consider whether it's aligned with your values.

Why we picked these: Our audience concerned citizens, not AI companies or policymakers. So, public comment and voting is the best way they can meaningfully contribute. Our main insight synthesizes the upshots of each strand of our project,

and it ties it together under the umbrella observation that change is happening rapidly, more rapidly than we may be aware of.

Data storyboard: We've decided to structure our narrative around three parallel strands. We can design interactions such as buttons at the left- and right-hand-side of the screen to jump between strands. Each strand is internally linear. Strand 1 is the most comprehensible to a non-expert; strands 2 and 3 require slightly more hand-holding

Strand 1: Carbon emissions	Train carbon intensity 	Deployment carbon intensity 
Explanation	Training occurs before deployment, so the train visualization comes first	We add more nuance to the story here, faceting deployment costs by type of model (and potentially linking to the previous visualization via a brush-and-filter)
Strand 2: Cybersecurity		
Explanation	Cybersecurity tries to catch attacks. First we will look at how effective AI is as a defensive tool	Now we will look at the economics of this defensive tool, showing that it's bringing down the cost of cybersecurity

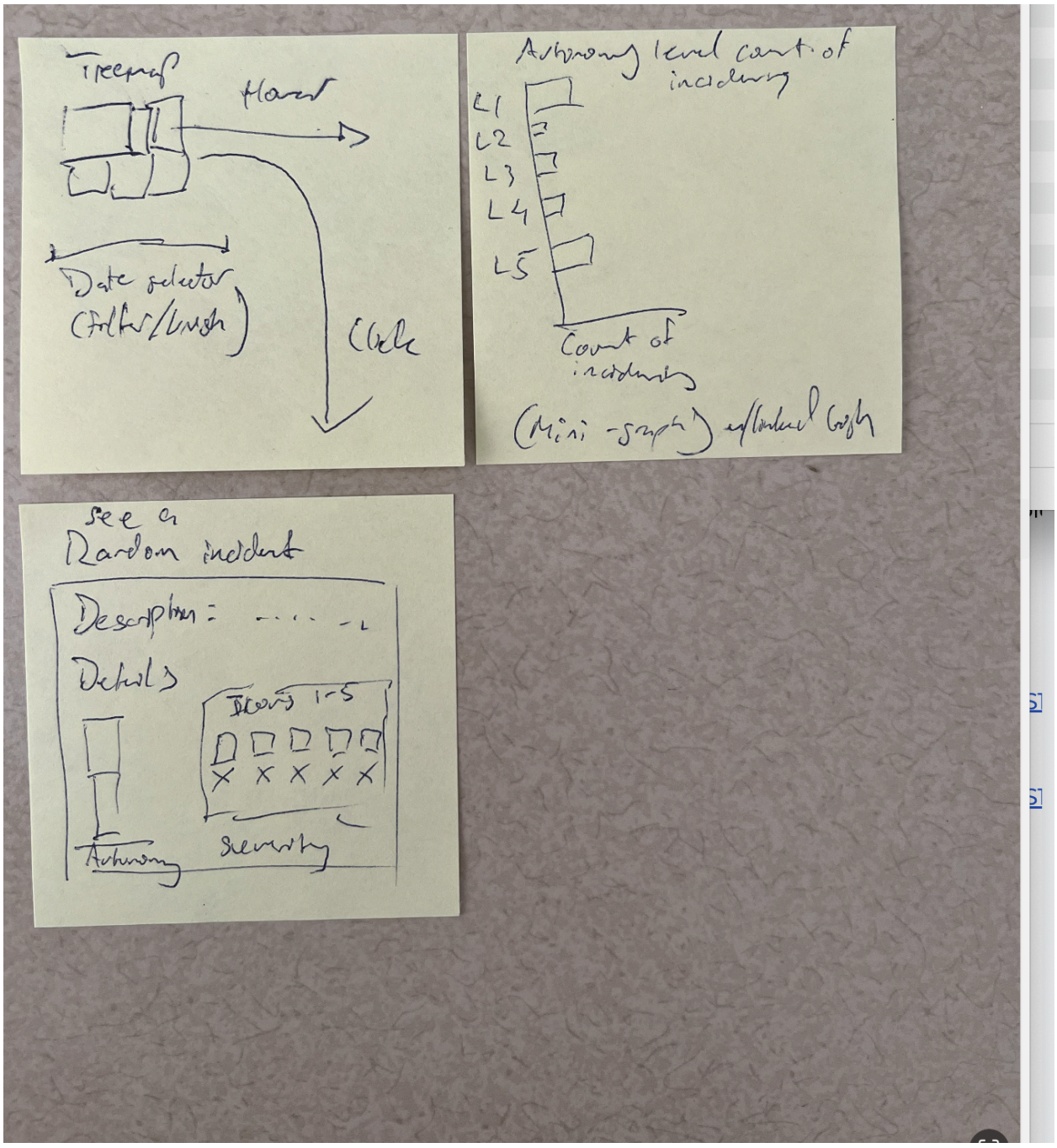
	(Depending on the data, could also flip this visualization to show the detections are decreasing, indicating offensive cyberattacks are getting better)	
Theme 3: Autonomy		 Hover detail into more about the incidents - e.g. what cognitive capacities they involved
Explanation	Situates model autonomy in context – the y axis changes during the interaction to represent different evaluations. Could pair with an explanation of what those evaluations mean (e.g., a sample problem)	Coming directly from AIID, this treemap allows users to explore a multidimensional dataset – sector x cognitive capacity – and look at the extent (or severity, maybe with a button to change the dimension) of incidents

Prototype V1

Drafts for other viz

(See above – our Tableau drafts are pretty detailed)

Interaction design (e.g., advanced viz)



We plan to implement selecting a random row on click to display further information from the AI Incidents Database – the detailed textual description, rating of how much autonomy was displayed, and severity of the incident per our categorization (represented with an icon)

This will let users explore the dataset more, and gain familiarity with what sort of incidents appear in each sector

We also will summarize characteristics of the sectoral cut of incidents on hover, with a pop-up mini-graph

The treemap will therefore be the entry point into a more detailed exploration of the data

Git instructions

1. Open Terminal >
2. Type cd RELATIVE/PATH/TO/FOLDER to get to the folder you want to use
3. Type git init to set up the repository
4. Then, add the latest
    a. ZIP method
        i. Ensure the .zip file is unpacked
        ii. Type git remote add origin https://github.com/xavierrobertsgaal/CS1710_Final_Project.git
    b. Clone method
        i. Type git clone https://github.com/xavierrobertsgaal/CS1710_Final_Project.git
5. Type git pull to confirm you have the latest