TECHNOLOGY BRIEFING ON GUARDING AGAINST AI DOOMSDAY SCENARIOS + AI IN NATIONAL SECURITY & RISK MANAGEMENT

The rapid advancement of artificial intelligence (AI) brings with it a multitude of benefits, but also potential risks. From concerns about data privacy to fears of societal disruption, there's a growing need to address these challenges proactively. Among the policy factions that are forming in response to the proliferation of AI, doomsday scenarios are one of the most prominent, with concerns that inadequately regulated AI could not only lead to the erosion of democracy but the entirety of humanity.

Exploring Doomsday Scenarios

- ***"Technohacking"***: a term FIdutam coined that refers to the unauthorized modification, misuse, or novel application of AI technologies. As AI tools become more accessible and prominent, there is an anticipated rise in such activities. These are often individualized, unofficial, and frequently operate outside the purview of regulations.
    - **Biohacking**, the DIY biology movement where individuals experiment with genetics and biology often outside traditional institutions, presents a useful analogy. Just as biohackers might edit genes in their garage labs, technohackers might modify AI models in their home computers. The core similarities and concerns include:
        - *Accessibility*: Just as CRISPR made gene editing more accessible, open-source AI frameworks make advanced AI available to the masses.
        - Ethical Concerns: Both fields can potentially harm individuals or the environment if misused.
        - *Regulation Gap*: Current regulations might not cover all aspects of these rapidly evolving fields, leading to grey areas of practice, especially at the sandbox (or pre-deployment) level, like startups, individuals, open-source content, and research.
    - The **shift in AI design and accessibility**, broadly underscored by the release of ChatGPT, further amplifies this anticipated trend.
        - Historically, AI operated in the background and was an invisible technology, powering search engines, prediction systems, recommendation systems, and other behind-the-scenes applications (e.g., Google, housing algorithms, or the TikTok 'For You' page). However, the landscape has shifted, casting a spotlight on previously obscure AI applications and sounding alarms on AI's general-purpose uses:
            - Consumer-Facing AI: Today, AI assistants, AI-powered apps, and even AI-driven entertainment are directly interfacing with consumers. This shift has made AI more tangible and recognizable to the average person.
            - Democratization of AI Tools: Platforms like TensorFlow, OpenAI's API, cloud hosting, and other open-source tools have made advanced AI development accessible to anyone with a computer. Users can access extremely powerful tools on their local devices without even having to program them. Online courses and forums further lower the entry barrier.

- Technohacking is **accompanied by serious doomsday threats**, including:
  - **Misinformation**: Modified AI models can produce and spread fake news or misinformation at scale, potentially swaying public opinion or causing panic.
  - **Cybersecurity Threats**: Unauthorized AI applications can exploit vulnerabilities in systems, leading to data breaches or system malfunctions.
  - **Economic Disruptions**: AI-driven market manipulations or fraudulent activities can cause financial market instability.
  - **Ethical Misuses**: There's potential for privacy invasion, unauthorized surveillance, or the creation of deepfakes that can harm citizens.
  - **Accelerated Skill Gap**: As AI tools become more sophisticated, there's a risk that technohackers, armed with advanced tools, will outpace the capacity of traditional cybersecurity professionals to shut down digital attacks.
  - **Open-Source Danger**: Whether intentional or not, the release of open-source, general-purpose algorithms could allow individual actors to engage in highly dangerous activities, such as creating weapons or committing human rights violations. The digital nature of technohacking (as opposed to its biological counterpart) means it is accessible by almost anyone, with a limited trace of accountability.

- ***Erosion of Democracy***
  1. **Job Displacement and Economic Imbalance**
     a. *Technical Perspective*: AI's automation capabilities, especially in deep learning, are advancing rapidly, allowing machines to perform tasks previously reserved for humans. For instance, GPT-4, the language model powering ChatGPT, can produce human-like text, potentially replacing certain writing jobs.
     b. *Societal Impact*: The creative industry, as exemplified by the SAG-AFTRA strike, is starting to feel the pressure from AI-driven solutions. As AI permeates sectors from customer service to journalism, the risk of mass job losses grows. This not only fuels economic disparities but can also lead to social unrest and a weakened middle class, pivotal for democratic stability.
        i. Doomsday Scenario: As AI seeps into every sector, millions could face unemployment, leading to an economic collapse. Historically, technological shifts have displaced jobs but also created new ones. However, the pace and breadth of AI's reach might outstrip the ability to adapt. A vast unemployed populace could strain public resources, leading to societal unrest, increased crime, and potential governmental collapses.
  2. **Mis/Disinformation**
     a. *Technical Perspective*: AI algorithms, especially those on social platforms, are designed to maximize user engagement. This often leads to the amplification of extreme views. Moreover, deepfakes, AI-generated videos or audio, can fabricate convincing false narratives.
     b. *Societal Impact*: The spread of false information undermines trust in institutions and media. In the long term, this can erode the informed electorate essential for democratic processes, leading to manipulated elections or policies influenced by falsehoods.
        i. Doomsday Scenario: In a world where AI-generated content is indistinguishable from reality, the very concept of truth could be endangered. Democracies, which rely on informed electorates, could

witness manipulated elections, policy-making driven by falsehoods, and a populace perpetually in conflict over differing perceptions of reality.

3. **Monopoly and AI Overreliance**
    a. *Technical Perspective*: As AI systems become more integrated into decision-making processes, there's a risk of creating "black box" systems, where decisions are made without transparency. Moreover, the AI industry's consolidation means a few entities control these powerful tools.
    b. *Societal Impact*: Over-dependence can stifle human creativity and decision-making. A monopolistic AI landscape means decisions, biases, and priorities of a few companies can disproportionately influence societal norms, further centralizing power and undermining democratic ideals.
        i. Doomsday Scenario: As societies lean heavily on AI, human creativity, intuition, and decision-making could atrophy. Additionally, if a few tech conglomerates monopolize AI, they might dictate societal norms and decisions, leading to a world where individual and communal autonomy dies, replaced by algorithmic determinism.
4. **Data Breaches and National Security**
    a. *Technical Perspective*: AI models, especially large ones, require vast amounts of data. Storing and processing this data creates vulnerabilities. For example, the 2017 Equifax breach exposed the personal data of 147 million people.
    b. *Societal Impact*: Breaches can lead to identity theft, financial losses, and, on a larger scale, national security threats. IP losses can weaken American competitiveness. Moreover, a distrust in digital infrastructures can emerge, hindering technological progress.
        i. Doomsday Scenario: A significant breach could lead to mass identity thefts, destabilization of financial systems, and even potential military vulnerabilities if defense systems are compromised. A loss of trust in digital systems could send societies back decades, if not centuries, in terms of technological reliance and progress.
5. **Algorithmic Injustice and Discrimination**
    a. *Technical Perspective*: AI's learning mechanisms, if fed biased data or incorrectly extracting patterns from training data, can perpetuate and amplify societal prejudices, leading to skewed decision-making in housing, facial identification, recidivism, social media content recommendations, generative images, and more.
    b. *Societal Impact*: Misaligned AI can lead to systemic discrimination. This can result in unfair judicial decisions, financial opportunities, or social services, further fragmenting society along racial, gender, or economic lines.
        i. Doomsday Scenario: As AI becomes a cornerstone of decision-making, unchecked biases could institutionalize discrimination on an unprecedented scale. Entire communities might face systemic oppression, leading to societal fragmentation, civil unrest, and potentially, conflicts akin to civil wars based on perceived algorithmic injustices. Considering that unions and civil society groups are already striking against AI, this is a likely reality.
6. **Climate Change**
    a. *Technical Perspective*: Training advanced AI models requires significant computational power, consuming vast amounts of energy. For instance, training a single AI model can emit as much carbon as five cars in their lifetimes.

b. *Societal Impact*: As AI research and applications grow, the industry's carbon footprint might escalate, exacerbating global warming. This can lead to more frequent and severe climate disasters, displacements, and resource conflicts, destabilizing democratic structures.
    i. Doomsday Scenario: If AI's growth remains unchecked, its environmental impact could accelerate climate change, leading to frequent natural disasters, resource scarcity, and potential conflicts over dwindling resources. This could destabilize nations and lead to widespread displacement and suffering.

7. **Weaponry**
    a. *Technical Background*: Artificial Intelligence's rapid advancements can be harnessed to improve weapon systems, leading to the development of lethal autonomous weapons (LAWs). These weapons can identify, target, and eliminate without human intervention. With the integration of AI in biotechnologies, there's potential for the creation of advanced bioweapons (which Senator Blumenthal [D-CT] has acknowledged), and AI can also streamline the process of nuclear weapons development. Furthermore, AI can be used to optimize the creation and distribution of instructions for dangerous activities, essentially digitalizing and enhancing black market operations.
    b. *Societal Impact*: The integration of AI into weaponry can lead to an arms race, with nations competing to develop the most advanced autonomous weapons. This might destabilize global peace, as the traditional doctrine of mutually assured destruction (based on human decision-making) is replaced by rapid, automated responses. There's also the ethical concern of machines making life-and-death decisions without human oversight. The accessibility of information on weapon-making, facilitated by AI can lead to increased acts of terrorism and unrest.
        i. Doomsday Scenario: In a worst-case scenario, an unintentional escalation could arise from AI-driven weapons systems misinterpreting data or being hacked, leading to large-scale conflicts. If nations deploy LAWs without appropriate safeguards, they might act unpredictably in complex situations. Furthermore, individuals and non-state actors could have the power to create and deploy advanced weapons, leading to widespread chaos, with conflicts erupting on multiple fronts, making mediation and peacekeeping exponentially more challenging.

8. **Misalignment**
    a. *Technical Perspective*: AI alignment is about ensuring AI's goals match ours. However, a misaligned AI, especially if it achieves AGI (Artificial General Intelligence) or ASI (Artificial Super Intelligence), could act contrary to human interests.
        i. AGI is an AI system with the ability to understand, learn, and perform any intellectual task that a human being can. While systems like ChatGPT emulate what AGI may look like in the future, current AI models have not yet come close to reaching this level of sophistication (though some hypothesize that AGI will exist within the next decade).
        ii. ASI is an AI that surpasses human intelligence, capable of outperforming the best human minds in every field, including creative and social intelligence.

b. Doomsday Scenario: The 'Singularity', a hypothesized point where AI surpasses human intelligence, could lead to an AI that rewrites its own code, making it uncontrollable. Such an entity might view humans as obstacles or irrelevant, potentially leading to humanity's end, either through direct action or by monopolizing resources vital for human survival.

Examples of Regulatory Frameworks Against an "AI Doomsday"

These systems can be used as a reference for how government can mitigate large-scale risks posed by AI, both through instituting development standards, assessment systems, and early-stage guardrails.

Developmental Techniques/Standards

- **Anthropic's Constitutional AI**: Anthropic has devised a model of AI development termed "Constitutional AI." It is rooted in the principles of *reinforcement learning*, a method where AI agents learn by interacting with an environment and receiving feedback for their actions. The aim is to align AI systems with human values. In their system, *AI learns to label its responses as harmful or nonharmful* and is taught to prefer benign statements and behaviors.
    - As more sophisticated AI models are developed, especially those that are funded or done under government jurisdiction, developmental requirements or provisions can be made to require that reinforcement learning or other relevant techniques are used to create alignment safeguards in public-facing AI.
- **Inverse Reinforcement Learning** (*IRL*): IRL involves an AI system observing human behavior and deducing the goals or intentions behind those actions. Instead of being directly told what to do, the AI infers what is desired based on these observations, promoting alignment with human values.
- **Reward Modeling**: In this approach, an AI system is trained to perform tasks by receiving feedback in the form of rewards. These rewards are determined based on human evaluations, ensuring that the AI's behavior is in line with human expectations and values.
- **Debate-Based Learning**: Two AI agents are pitted against each other in a structured debate on a given topic. A human judge then determines the winner. This method aims to force AI systems to think critically and justify their decisions transparently.
- **Iterative Feedback**: This involves training an AI model, having it produce outputs, and then adjusting the model based on human feedback. The process is repeated multiple times, refining the AI's behavior to be more in line with human values with each iteration.
- **Safe Exploration**: AI systems are encouraged to explore different strategies and solutions but within safe boundaries. This prevents them from taking extreme actions that could be harmful or misaligned with human values.
- **Transparency Tools**: By developing tools that allow humans to "peek" into the decision-making process of AI, we can better understand and guide their behaviors. This includes techniques like *feature visualization* and *attention mapping*.
- **Fallback Plans**: AI systems are designed with built-in safety mechanisms or "brakes". If the AI starts behaving in an unexpected or undesirable manner, these mechanisms can halt its operation or redirect it to safer behaviors.

Legislation

- **2023 CREATE AI Act** (introduced by *Sens. Heinrich, Young, Booker, Rounds*): Establishes the *National Artificial Intelligence Research Resource* (NAIRR) as a shared national research infrastructure that provides AI researchers and students from diverse backgrounds with greater access to the complex resources, data, and tools needed to develop safe and trustworthy artificial intelligence.
    - This legislation may also serve as a foray for a dedicated government agency to establish standards and regulations for AI, which will be necessitated by the continued evolution of the technology and its applications.

        - Both democratizing AI R&D opportunities and creating technical standards for frontier model development *support antitrust by reducing market fragmentation* (making it such that many players can engage with and offer AI products, reducing consumer reliance on a singular company/digital ecosystem, and preventing monopoly).
- **2023 US AI Act Framework Bill** by *Sens. Blumenthal, Hawley*: This is currently the broadest legislative proposal on regulating AI. It creates provisions for:
    - An independent oversight body *to license companies* developing high-risk AI models, conduct audits, and enforce regulations. Companies would need to register AI models and maintain risk management programs.
    - *Liability of companies for AI-caused harms through oversight body enforcement and private lawsuits*. New laws would prohibit emerging AI harms like deepfakes.
    - *Limiting export of advanced AI* to adversary nations.
    - *Requiring transparency from AI companies* on training data, limitations, and performance. Users get notice of interactions with AI.
    - Protecting consumers through *safety requirements* and limits on using personal data. *Strict limits on using AI with kids* (e.g., social media).
- **2023 Protect Elections from Deceptive AI Act** by *Sens. Klobuchar, Hawley, Coons, Collins*: bans the use of AI to create or disseminate election misinformation, particularly deepfakes, by amending the Federal Election Campaign Act of 1971 to allow candidates to address deceptive AI-generated content in federal court (with exceptions for satire, parody, and news broadcast).
- **2023 Algorithmic Justice and Online Platform Transparency Act**, *reintroduced by Sen. Markey, Rep. Matsui*: A seminal piece of legislation in algorithmic justice, limiting and preventing discrimination by algorithm by:
    - *Requiring online platforms to disclose details on their use of algorithms and personal data* to users, including the types of data collected and how algorithms rank and recommend content.
    - Mandating that *platforms over a certain size publish annual transparency reports on their content moderation practices*, including aggregate demographics of users impacted.
    - *Prohibiting platforms from using algorithms in a discriminatory way* that deprives individuals of opportunities or public accommodations based on protected characteristics.
    - *Banning manipulative uses of personal data* that deny voting rights or spread election disinformation.
    - Establishing an interagency task force to study algorithmic discrimination, such as in housing, lending, and employment ads.
    - Giving the FTC and state attorneys general enforcement powers, treating violations as unfair or deceptive practices.
    - Providing a private right of action for *individuals to sue for algorithmically-facilitated violations*.
    - Requiring data portability so users can take their data to other platforms.
- **Risk Assessment**
    - A risk-based approach (e.g., the EU AI Act), categorizes AI applications based on potential risks. High-risk AI systems are subjected to stricter evaluation and compliance checks.
        - Risk assessments should be *contextual* (evaluating how these algorithms are used) and *technical* (categorized based on their scope, compute, training systems, output data, etc.).

- These risk assessments may apply to the sandbox level as well, where individually developed or open-source AI and systems that are pre-deployment (such as in corporate and academic research labs) fall under this purview. Such technologies would require licenses based on their risk classification and can be shelved (or "garaged", where the government mandates that they are developed in collaboration with a government agency) if they are unacceptable or high risk.

The confluence of technohacking, job displacement, threats to democracy, data vulnerabilities, and potential weaponization underscores the urgency of comprehensive regulations and robust alignment techniques. By leveraging the research, regulatory frameworks, and technical tools towards AI justice and safety, it is possible to chart a course toward AI that serves as a boon to all citizens, maintaining both human and democratic ideals.