The Hydra’s First Head - Ground Truth

Origins of the threat

Conflicts over dwindling natural resources, climate-induced mass immigration, and increasingly frequent pandemics have steadily eroded international relations. Governments and citizens have become highly vulnerable to populist and nationalist rhetoric, and international warfare appears increasingly probable.

An agency of unknown affiliation is capitalizing on this situation, aiming to destabilize global order. They use powerful machine learning algorithms to identify vulnerable but tech-competent citizens and radicalize them to develop and release powerful disinformation AI programs. Using AI to automate and progressively improve this process enables massive scale with a high tolerance for failure. Successful attacks show high variability in algorithm design and function, creating a many-headed ‘hydra’ that is difficult to effectively combat. Though targets may be quickly identified following an attack, the agency itself remains undiscovered.

The scenario describes the initial stage of the first majorly successful disinformation campaign, using AI programs developed by Arno Jackson, a US citizen unknowingly manipulated by this agency. The campaign begins with the release of a highly convincing deepfake video of the US President advocating the use of nuclear weapons against China. This video and countless others on similar themes are rapidly distributed online to international audiences, leading to escalation in global tensions and growing possibility of nuclear war.

The China-Taiwan situation

A month prior to this scenario, The People's Republic of China staged a comprehensive naval blockade of Taiwan, following decades of pressure on the island to unify with mainland China. The blockade created global economic disruption by halting Taiwan’s semiconductor exports. While China claims it does not seek conflict with the US or others, many saw it as a move to consolidate the nation’s influence and test global power dynamics. Despite this, most countries have refrained from outright condemnation of the blockade.

Taiwan’s government has called on the US for military assistance, a request which has so far gone unanswered, partly due to pressure from NATO allies to remain neutral. The US finds itself in a Catch-22 situation: if China invades and conquers Taiwan, the US’s status as the world’s most powerful nation is seriously challenged; on the other hand, the US’s defense of Taiwan would risk all-out war.

Though physical warfare has so far been avoided, the digital world has become an international battleground plagued by frequent cyberattacks on nations’ military databases, health systems, government records, and critical infrastructure. Coupled with growing concerns about the potential use of nuclear weapons, some say ‘The Second Cold War’ is underway.

Taiwan is the world’s leading creator of computer chips, especially advanced semiconductors used in AI applications. Despite the US and other nations investing in domestic foundries in recent years to reduce dependency on Taiwan, ever-rising global demand for chips has sustained Taiwan’s position as the most critical supplier. The blockade created an immediate crisis for the computer chip industry, as trade with other nations became impossible and Taiwanese foundries shut down their operations. Chip demand from many nations’ militaries also skyrocketed following the blockade, as they fast-tracked the development and rollout of AI-powered weapons and defense systems, including smart drones, cyber-weapons, and satellite target recognition systems, in preparation for potentially imminent conflict. This further exaggerated the supply issue.

Deepfake developments

AI technologies have continued to grow in application, sophistication and accessibility, bringing both significant benefits and risks to society. Deepfake technologies have become increasingly powerful, with limited legitimate uses including recreation of damaged photo or video footage or crime reconstruction. Malicious applications, especially to create and spread false information, are abundant.

Deepfake detection software remains underdeveloped and underutilized. Shareholders of social media and news platforms have observed little correlation between engagement with online media and truthfulness of said media, so are reluctant to prioritize the high investment required to develop counter-algorithms to detect misinformation, given relatively few perceived returns.

US misinformation policy

US governmental policy continues to lag behind technological advances. The scenario described takes place early in the term of the new Presidential Administration, at a time of deep public mistrust in government. A main policy platform of the incumbent Presidency is a clamp-down on the proliferation of false information. Following election victory, the addition of the Presidential Truthful Information Advisor to the Presidential Cabinet was announced, for which MIT alumnus Dr Harry Shah was appointed and quickly dubbed ‘Truth Guru’.

A more comprehensive effort to address the issue was the proposal of the Misinformation, Disinformation, and Malinformation Bill, which would place the onus on online platforms to detect and counter false information or face strong penalties. Progress on the Bill’s passage has stalled due to a sizable portion of Congress arguing that it encourages government overreach.

Response

The US government recognizes the President deepfake as a threat to national security and quickly assembles a reactive taskforce, overseen by Dr Harry Shah. They realize that current AI-detection tools are insufficient to tackle the software used to create the video, and so set about fast-tracking the development of a new detection tool. Through collaboration across government departments and with MIT’s machine learning team, they are able to develop the tool within a matter of days and effectively combat the next-gen deepfakes.

However, the taskforce acknowledges that the tool will not necessarily counter future attacks, and that a strategic, proactive approach to address increasingly-sophisticated disinformation attacks is desperately needed.

Though online for less than a week, the deepfake videos were seen by millions and still caused significant harm. At the societal level, the US experienced increased rates of hate crimes against Asians, which persisted after the videos were debunked. Similar trends occurred in China, including an arson attack on an international school. At the governmental level, diplomatic communications between the US and China briefly ceased, and both nations readied their nuclear arsenals. Despite de-escalation following acknowledgement that the videos were fake, the realization that both nations appeared willing to follow through on nuclear threats led to long-lasting damage to international relations.