Issue Report: Impact of AI on Disinformation Campaigns

As of 20 June 2023

INFER recently launched a set of questions that informed a talk at the Phoenix Challenge in late June about the impact of AI on influence operations and disinformation trends. INFER's forecasts were presented during the talk and compared to forecasts from event attendees, including government defense delegations, executives, experts and policymakers from a host of allied countries and institutions.

Will a country ban or take regulatory actions that ultimately block access to OpenAI's models, between 1 Jun 2023 and 31 Oct 2023, inclusive?

Crowd Forecast

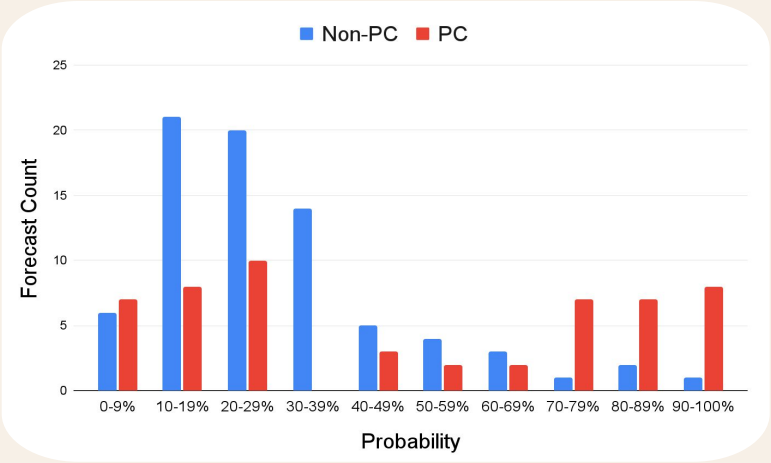
36% chance

131 forecasters

Phoenix Challenge

54% chance

54 forecasters



See detailed forecast rationales • See consensus trend and crowd profile

Before 1 June 2024, will Facebook, WhatsApp, Messenger, or Twitter announce that they are labeling posts as potentially written by AI?

Crowd Forecast

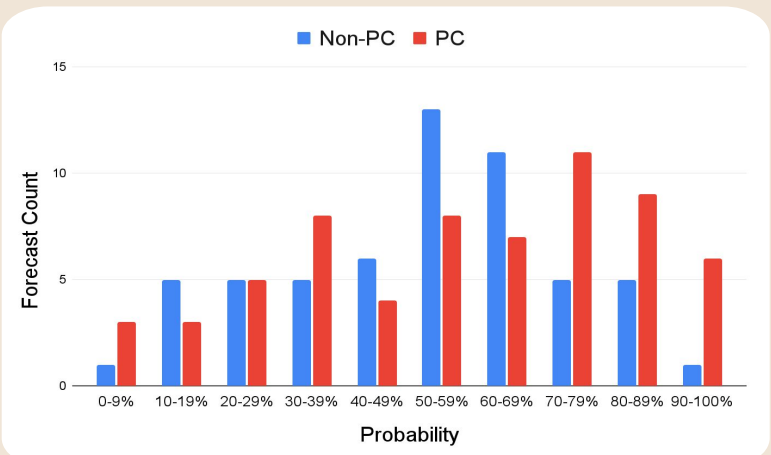
52% chance

120 forecasters

Phoenix Challenge

58% chance

67 forecasters



See detailed forecast rationales • See consensus trend and crowd profile

How many people will have signed up for World ID on 1 September 2023? (Answer: 4 million+)

Crowd Forecast

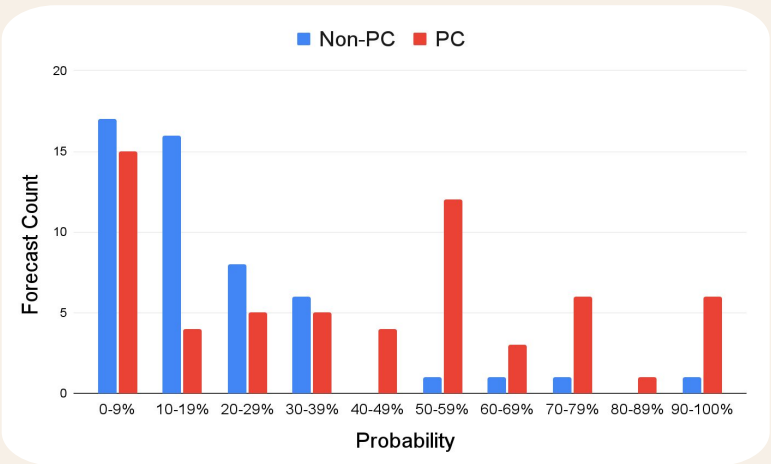
14% chance

111 forecasters

Phoenix Challenge

42% chance

64 forecasters



See detailed forecast rationales • See consensus trend and crowd profile

Will YouTube, Facebook, Instagram, or Twitter enable digital provenance (e.g., C2PA) on photos or videos in 2023?

Crowd Forecast

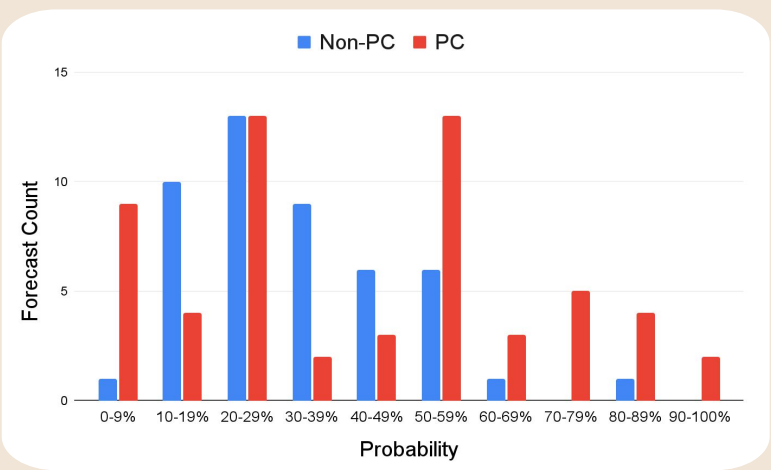
29% chance

105 forecasters

Phoenix Challenge

42% chance

58 forecasters



See detailed forecast rationales • See consensus trend and crowd profile

Will any of Meta's 2023 threat disruption reports indicate that a LLM may have been used to conduct an influence operation?

Crowd Forecast

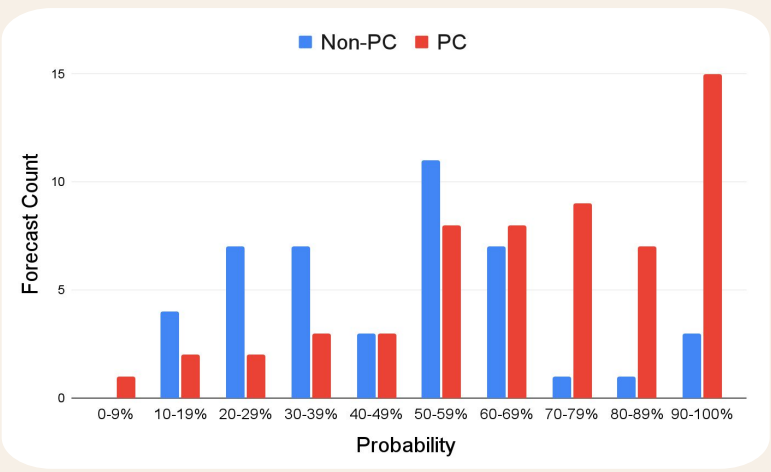
54% chance

102 forecasters

Phoenix Challenge

70% chance

58 forecasters



See detailed forecast rationales • See consensus trend and crowd profile

Will the New York Times, CBC, or BBC announce that they will only publish photos or videos containing digital provenance (e.g., C2PA) by 31 May 2024?

Crowd Forecast

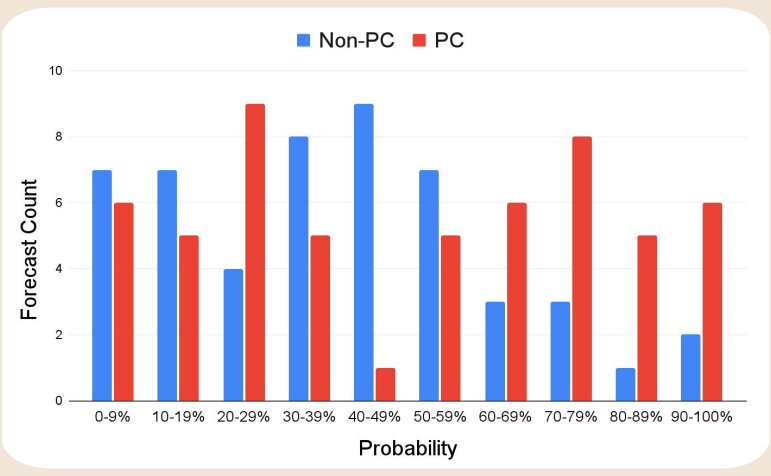
36% chance

106 forecasters

Phoenix Challenge

53% chance

56 forecasters



See detailed forecast rationales • See consensus trend and crowd profile

APPENDIX: Forecaster Rationale Summaries

This section presents a high-level summary of forecasters' rationales for each question in the report. Rationales can be found in full by clicking “See detailed forecast rationales”, and a list of sources linked within rationales can be found by clicking “See source links”. The data in this report is as of 20 June 2023.

A: Forecast Questions

Will a country (currently supported by OpenAI) ban or take regulatory actions that ultimately block access to OpenAI's models, between 1 June 2023 and 31 October 2023, inclusive?

TOTAL FORECASTS: 162 | PC FORECASTS ONLY: 56

Consensus Forecast By Audience	
36% chance	Crowd Forecast
54% chance	Phoenix Challenge (PC) Participant Forecast

Summary of Rationales By Audience

See detailed forecast rationales | See source links

INFER Crowd (non-PC)	Phoenix Challenge Participants
Arguments why a country WILL block access:	
• **Privacy concerns:** Several countries, especially in the EU, are investigating OpenAI's data handling practices. • **Regulatory concerns:** Regulators may want to ensure AI models comply with laws around data privacy, security, bias and unfairness before they become widely adopted. E.g., the EU is working on AI-specific regulations that could apply to OpenAI. • **Authoritarian regimes:** Though most authoritarian regimes are already not supported by OpenAI, there may be some not on the list wanting to ban OpenAI to control information and dissent.	• **Privacy concerns:** [Same point as INFER audience] • **Regulatory concerns:** [Same point as INFER audience] • **Job loss concerns:** Some forecasters argued that countries may ban OpenAI to protect jobs, especially if models like ChatGPT become widely used.
Arguments why a country WILL NOT block access:	
• **Slow policy process:** Unlikely most governments can implement a ban within 5 months given typical legislation and regulation often moves slowly. • **Benefits of technology:** New technologies like AI drive innovation and economic growth. A ban may stifle AI progress and put them at a competitive disadvantage. • **Lack of incidents:** There is little evidence to suggest that OpenAI's models currently warrant major regulatory action or pose existential threats.	• **Slow policy process:** [Same point as INFER audience] • **Benefits of technology:** [Same point as INFER audience] • **Limited impact:** Even if a country bans ChatGPT, many other groups are working on similar technologies. Countries might not ban OpenAI's models if they believe that it would not ultimately stop progress in AI.

Before 1 June 2024, will Facebook, WhatsApp, Messenger, or Twitter announce that they are labeling posts as potentially written by AI?

TOTAL FORECASTS: 133 | PC FORECASTS ONLY: 66

Consensus Forecast By Audience	
52% chance	Crowd Forecast
60% chance	Phoenix Challenge (PC) Participant Forecast

Summary of Rationales By Audience

See detailed forecast rationales | See source links

INFER Crowd (non-PC)	**Phoenix Challenge Participants**
Arguments why a company WILL label posts:	
• **External pressure:** There is increasing public and regulatory pressure for transparency about AI use and to curb AI-enabled disinformation. Labeling posts could help address these concerns and pressures. Several forecasts point to interests from the EU and US governments. • **Detection technology improving:** AI detection technology, while imperfect, is improving and some tools already exist. Even if not fully accurate, platforms may view some level of labeling as better than none. Some forecasts argue the technology is "good enough." • **U.S. presidential election:** 2024 elections will increase pressure to address AI disinformation.	• **Regulatory pressure:** Many forecasters argue that regulatory bodies like governments (especially the EU) and Congress will pressure social media companies to increase transparency around AI-generated content. Companies may label posts to preempt legislation and regulation. • **Public demand for transparency:** Public interest groups and users will demand more transparency into AI-generated content, pushing companies to label posts. • **AI detection improving:** AI detection technology is improving quickly and will reach a level of accuracy in the next year that enables companies to label posts with a high degree of confidence.
Arguments why a company WILL NOT label posts:	
• **Little business incentives:** There is little incentive or motivation for the companies to label posts. Their main goal is to maximize engagement and profits, and labeling may undermine that. • **Limited detection technology:** AI detection technology is still limited and error-prone, leading to many false positives and negatives, which could frustrate users and create backlash. May be technically difficult and expensive to implement labeling at scale across platforms.	• **Little business incentives:** [Same point as INFER audience] • **Limited detection technology:** [Same point as INFER audience] • **Legal liability concerns:** A few forecasters argue that labeling posts as AI-generated could make companies legally liable for the content, creating an incentive for them not to label.

• **Lack of motivation for messaging apps:** Messenger and WhatsApp are focused on messaging, not public posting, so they may be even less motivated or able to label AI content.	

How many people will have signed up for World ID on 1 September 2023?

TOTAL FORECASTS: 131 | PC FORECASTS ONLY: 63

Consensus Forecast By Audience	
15% chance of 4 million or more sign ups	Crowd Forecast
43% chance of 4 million or more sign ups	Phoenix Challenge (PC) Participant Forecast

Summary of Rationales By Audience See detailed forecast rationales | See source links	
INFER Crowd (non-PC)	**Phoenix Challenge Participants**
Arguments why World ID will have FEWER than 4 million sign ups by 1 September 2023:	
• **Trust and privacy concerns:** New technologies often face trust barriers, esp. those that deal with the digitalization of personal data. Current messaging fails to address privacy concerns in a compelling and easily understood way. • **Tech not widely adopted:** World ID and the underlying blockchain tech are still niche and not widely adopted by the general public. Without mainstream adoption and use cases, people may be more reluctant to sign up. • **Crypto hype declining:** Interest in crypto has declined recently, so people may be less interested in a new crypto-related service.	• **Trust and privacy concerns:** [Same point as INFER audience] • **Not well known:** Many of the forecasters themselves are largely unfamiliar with World ID, and they express doubt that if they themselves have not heard of it or do not fully understand what it is, it's unlikely to be well known to most ordinary people and mainstream audiences. • **Unclear value proposition:** Forecasters see few real-world use cases demonstrating value and think the current messaging lacks broad appeal.
Arguments why World ID will have 4 million or MORE sign ups by 1 September 2023:	

• **Appeals to enthusiasts:** Enthusiasm from groups like crypto and tech communities generates early traction and could drive strong initial growth. • **Steady growth:** World ID's growth is a modestly positive signal of gradual progress, but forecasters still question whether there has been enough fundamental change or mainstream interest growth to fuel exponential adoption in the near term. • **Incentives to join:** Crypto tokens and other financial incentives may motivate some niche groups, but stronger incentives, use cases, and benefits for mainstream audiences would be needed to drive mass adoption.	• **Appeals to enthusiasts:** [Same point as INFER audience] • **Steady growth:** [Same point as INFER audience] • **Worldwide potential:** A large potential target market could drive high adoption.

Will YouTube, Facebook, Instagram, or Twitter enable digital provenance (e.g., C2PA) on photos or videos in 2023?

TOTAL FORECASTS: 115 | PC FORECASTS ONLY: 60

Consensus Forecast By Audience	
29% chance	Crowd Forecast
42% chance	Phoenix Challenge (PC) Participant Forecast

Summary of Rationales By Audience See detailed forecast rationales | See source links	
INFER Crowd (non-PC)	**Phoenix Challenge Participants**
Arguments why a company WILL label posts:	
• **Easy to implement:** Digital provenance may be a relatively easy feature for platforms to implement to show they are taking action. • **Pressure to address misinformation:** Advertisers may demand platforms enable provenance to avoid associating brands with misinformation. It may also satisfy regulators and lawmakers calling for platforms to address misinformation. • **Desired transparency as AI improves:** As AI's capabilities around generating media	• **Easy to implement:** [Same point as INFER audience] • **Twitter interest:** Twitter CEO Elon Musk appears particularly interested in introducing new information-verification tools on the site. • **Improve trust:** Enabling provenance can help build trust with users and address misinformation.

increase, platforms may enable digital provenance to give users more transparency.	
Arguments why a company <u>WILL NOT</u> label posts:	
• **Financial implications:** Implementing provenance technology and infrastructure can be costly for platforms, and there is little financial incentive. • **Technical challenges:** Provenance technology like C2PA is still new and developing, not yet ready for large-scale implementation. There is also a lack of C2PA-compliant equipment which limits the usefulness of implementing the standard. • **Competing priorities:** Platforms have other priorities, like labeling AI-generated content, that may take precedence over enabling provenance.	• **Financial implications:** [Same point as INFER audience] • **Slow adoption:** Digital provenance may be slow to gain mainstream adoption by users, reducing incentives for platforms to enable it. • **Competition with provenance proponents:** Google and Facebook may not want to support a standard created by a competitor like Microsoft.

Will any of Meta's 2023 threat disruption reports indicate that a LLM may have been used to conduct an influence operation?

TOTAL FORECASTS: 107 | PC FORECASTS ONLY: 60

Consensus Forecast By Audience	
54% chance	Crowd Forecast
70% chance	Phoenix Challenge (PC) Participant Forecast

Summary of Rationales By Audience See detailed forecast rationales | See source links	
INFER Crowd (non-PC)	**Phoenix Challenge Participants**
Arguments why a threat disruption report <u>WILL</u> indicate a LLM was used to conduct an influence operation:	

• **LLMs easily accessible:** Because LLMs have become widely available and easy to use, malicious actors will likely deploy them for disinformation. • **Clear threat of weaponization:** The capabilities and incentives for using LLMs to spread propaganda and manipulate public opinion are clear, suggesting that their misuse is unavoidable if not already ongoing. • **Automated content generation:** LLMs can quickly and automatically produce a high volume of synthetic media, comments, posts and accounts, enabling coordinated inauthentic behavior at a pace and scale that exceeds human efforts.	• **Adversaries' investment:** Malicious actors, including state adversaries, have spent significantly on advancing their AI capabilities. LLMs would be an obvious tool for them to deploy in psychological and information warfare. • **Low-cost and scalable:** LLMs can conduct and amplify influence campaigns in a highly automated, high-volume fashion at relatively low cost. • **Meta likely to report:** If they found evidence of a LLM being actively used to manipulate their platforms, they would likely disclose it to demonstrate the importance and impact of their efforts to counter the use of AI to spread misinformation.
Arguments why a threat disruption report <u>WILL NOT</u> indicate a LLM was used to conduct an influence operation:	
• **Difficult to detect:** Determining whether a LLM was involved in an influence operation would be technically challenging and require substantial evidence. • **Incentives not to disclose by Meta:** Meta may choose not to report the use of LLMs for corporate or political reasons. • **Lack of evidence of real-world misuse:** There is little evidence so far that LLMs have actually been used for large-scale influence operations or other malicious purposes.	• **Difficult to detect:** [Same point as INFER audience] • **Incentives not to disclose by Meta:** [Same point as INFER audience] • **Still too early:** The use of LLMs in influence operations may be inevitable, but the operations may not yet be sophisticated enough or at a large enough scale for Meta to state that LLMs were used.

Will the New York Times, CBC, or BBC announce that they will only publish photos or videos containing digital provenance (e.g., C2PA) by 31 May 2024?

TOTAL FORECASTS: 116 | PC FORECASTS ONLY: 58

Consensus Forecast By Audience	
36% chance	Crowd Forecast
53% chance	Phoenix Challenge (PC) Participant Forecast

Summary of Rationales By Audience See detailed forecast rationales | See source links	
INFER Crowd (non-PC)	**Phoenix Challenge Participants**
Arguments why a company WILL label posts:	
• **Adoption by tech firms:** Some tech companies like Microsoft and Adobe are already using the C2PA standard, so others may follow. • **Regulatory pressure:** Some lawmakers want tech companies to address AI-generated images, so enabling C2PA may appease regulators. • **User demand:** If C2PA becomes popular and users start demanding it, the platforms may enable it to stay ahead.	• **Adoption by tech firms:** [Same point as INFER audience] • **BBC's involvement:** The BBC helped develop C2PA, so they may announce using it. • **Improves trust:** Media companies rely on trust and credibility, so they have incentive to adopt digital provenance to help address manipulated images.
Arguments why a company WILL NOT label posts:	
• **Technical challenges:** The technology for digital provenance is not ready or widespread enough for platforms to implement at scale. • **Financial costs:** Implementation would be too costly and restrictive for media organizations, and there is little incentive or demand for them to adopt C2PA. • **Regulation too slow:** Regulation and policy changes happen too slowly to require platforms to enable digital provenance by the end of 2023.	• **Technical challenges:** [Same point as INFER audience] • **Financial costs:** [Same point as INFER audience] • **Importance of publishing speed:** The desire to publish news quickly outweighs the need to verify authenticity, so they won't commit to only using provenance-verified content.

B. The Forecasters

The forecasters who have participated thus far in these six questions have the following profile:

- 162 total forecasters
- 31% are “INFER Pros” - participants selected for the Pro Program due to their proven accuracy track record on INFER or with a similar forecasting site/program
- 40% are participants in the Phoenix Challenge

Demographics

Country	**Ratio**
USA	40%
Canada, UK, Australia, New Zealand	9%
Europe	32%
Latin America, Caribbean	10%
Asia	10%

C. Report Methodology

"Crowd forecasts" refer to the consensus forecasts generated by everyone who forecasted on a given question. “Phoenix Challenge” or "Phoenix Challenge (PC) Participant Forecasts" refer to the consensus forecasts generated only by attendees of the Phoenix Challenge conference.

Histograms on the first page of the report compare the distribution of forecasts made by attendees of the Phoenix Challenge conference (i.e., "PC") to all other forecasters (i.e., "Non-PC") on a given question.

Rationales of PC and non-PC INFER forecasters have been summarized by Claude, an AI assistant tool created by Anthropic. To build the rationale summaries presented in this report, we provided forecasts (probabilities and narrative rationales) to Claude to summarize into bulleted lists of arguments. We then manually edited the bulleted summaries for accuracy and readability.

Each question in the report also includes links to the crowd forecasts, rationales, and source links used by forecasters.