Answer: Advanced automated forecasting better enables political leaders to avoid precarious moments that could spark a large-scale conflict.

5. **Result Fragility.** Do the findings rest on strong theoretical assumptions; are they not demonstrated using leading-edge tasks or models; or are the findings highly sensitive to hyperparameters? ☐
6. **Problem Difficulty.** Is it implausible that any practical system could ever markedly outperform humans at this task? ☐
7. **Human Unreliability.** Does this approach strongly depend on handcrafted features, expert supervision, or human reliability? ☐
8. **Competitive Pressures.** Does work towards this approach strongly trade off against raw intelligence, other general capabilities, or economic utility? ☐

C.2 Safety-Capabilities Balance

In this section, please analyze how this work relates to general capabilities and how it affects the balance between safety and hazards from general capabilities.

9. **Overview.** How does this improve safety more than it improves general capabilities?
   Answer: While this line of work reduces systemic risk factors and can improve institutional decision making, making AI systems better at forecasting could potentially improve general capabilities. Its relation to general capabilities is currently unclear. In humans, at the extremes, IQ is hardly predictive of forecasting ability, suggesting forecasting of near-term geopolitical events is a specific and not general skill. Likewise, work in this space could focus on engineering better forecasting systems rather than improving general representations, so as to avoid capabilities externalities; this is potentially a more robust strategy for avoiding capabilities externalities. If it turns out that capabilities externalities are difficult to avoid even while simply engineering better forecasting systems, we would suggest that safety researchers stop working on this problem.
10. **Red Teaming.** What is a way in which this hastens general capabilities or the onset of x-risks?
    Answer: Making AI systems better at forecasting could also improve general capabilities or at least the raw power of AI systems. As Yann LeCun reminds us, "prediction is the essence of intelligence."
11. **General Tasks.** Does this work advance progress on tasks that have been previously considered the subject of usual capabilities research? ☐
12. **General Goals.** Does this improve or facilitate research towards general prediction, classification, state estimation, efficiency, scalability, generation, data compression, executing clear instructions, helpfulness, informativeness, reasoning, planning, researching, optimization, (self-)supervised learning, sequential decision making, recursive self-improvement, open-ended goals, models accessing the internet, or similar capabilities? ☒
13. **Correlation With General Aptitude.** Is the analyzed capability known to be highly predicted by general cognitive ability or educational attainment? ☐
14. **Safety via Capabilities.** Does this advance safety along with, or as a consequence of, advancing other capabilities or the study of AI? ☒

C.3 Elaborations and Other Considerations

15. **Other.** What clarifications or uncertainties about this work and x-risk are worth mentioning?
    Answer: Regarding Q7, while human forecasters are important for building a training set with rich annotations, the actual human forecasts are unnecessary, as technically only the resolutions are needed. Additionally, the end goal is to create automated forecasting systems that do not depend on human reliability. Eventually, these systems could become much faster and more reliable than human forecasters.

    Regarding Q12, this work facilitates research towards general prediction of future events and consequently toward improved planning. However, we expect the kinds of predictions improved by forecasting research to be especially relevant for reducing x-risk. For example, improved institutional decision making surrounding geopolitical events could reduce the risk of global conflicts leading to the weaponization of strong AI.

Regarding Q13, IQ is predictive of forecasting ability in humans, not overwhelmingly so (Mellers et al., 2015). Moreover, its correlation is especially weak at extremes. Likewise, forecasting skills for near-term geopolitical events are partly learnable, further suggesting a separation from general cognitive ability.

Regarding Q14, while the relationship between general capabilities and research on forecasting near-term geopolitical events is currently unclear, this research does advance the study of narrow AI systems.

Finally, we would like to discuss limitations and potential hazards of relying on ML for forecasting near-term geopolitical events.

(a) Forecasting is best used for refining understanding rather than for anticipating the future more generally. Forecasters are demonstrated to be useful for optimizing probabilities for somewhat likely events (e.g., events with probabilities between, say, 5% and 95%). What is more important are tools that unearth important considerations that were implicitly assigned negligible probabilities or wrongly treated by humans as misinformation or worth ignoring. These considerations are often not forecasted and are not thought worth asking; implicitly, such events could the thought to be assigned low probabilities (e.g., say 10^{-7}), while some people argue that these considerations are more likely than others believe (e.g., say 10^{-1}). The information value provided from putting ignored considerations on our radar is substantial, in fact, orders of magnitude greater than the information gained by refining probabilities by a few percent. Forecasting competitions are about refining estimates of known unknowns–questions already on our radar–but what is better for risk reduction is confronting unknown unknowns, finding considerations to put on our radar, and reducing *exposure* to inchoate potential risks. For this reason, Hendrycks et al. (2021c) suggest tools that improve brainstorming and suggesting considerations.

(b) Forecasting is not necessarily a suitable tool for addressing tail risks. Taleb and Tetlock (2013) remind us that "No one has yet figured out how to design a forecasting tournament to assess the accuracy of probability judgments that range between .00000001% and 1%—and if someone ever did, it is unlikely that anyone would have the patience–or lifespan–to run the forecasting tournament for the necessary stretches of time (requiring us to think not just in terms of decades, centuries and millennia)." Taleb and Tetlock (2013) further remind us that it is unjustified to use forecasting tools for revolutions, market crashes, venture capital, or other winner-take-all domains. Furthermore they note that framing questions about tail risks as "a binary question is dangerous because it masks exponentially escalating tail risks." Consequently, "improving short-run probability judgments" and "contingency planning for systemic [tail] risks" are "complementary" and separate (Tetlock et al., 2022). Indeed, superforecasters usually anchor in outside view (Tetlock and Gardner, 2016), which neglects systemic risks. In environments with tail events, it is not how often one is correct that matters but rather how large one's cumulative errors are; current forecasting metrics do not sufficiently penalize forecasters that ignore tail risks nor do they greatly reward prescience about Black Swans.

(c) Forecasting tools could lead to risky behavior. For example, forecasting systems may induce inaction. If forecasts are uncertain, leaders may argue that "we should not make a decision before we have a reliable forecast" so we should "sit tight and assess." This is sometimes referred to as the delay fallacy, namely "if we wait we will know more about X, hence no decision about X should be made now" (Hansson, 2004). However, it is often cheaper to prevent risks or reduce exposure to risks, as "an existential risk needs to be killed in the egg, when it is still cheap to do so" (Taleb et al., 2020). Waiting until all the relevant information arrives is often waiting until it is too late.

Furthermore, humans are known to misinterpret probabilities (Vodrahalli et al., 2022). Systems that assign an event 3% probability may lead decision-makers to assume the event will not happen. Automation bias may mean forecasting systems induce users to have a gain in confidence that is greater than their gain in knowledge. Risk compensation suggests this could result in riskier actions (Hedlund, 2000). Furthermore, forecasts are often not provided with reverse psychology in mind. However, a forecasting system that forecasts a low risk can lead users to act as though there is no risk and increase risky behavior, which increases systemic risk.