FFI-RAPPORT

21/00735

Hvordan forbedre treffsikkerheten til prediksjoner av internasjonal politikk?

— en litteraturgjennomgang

Alexander William Beadle

Hvordan forbedre treffsikkerheten til prediksjoner av internasjonal politikk?
– en litteraturgjennomgang

Alexander William Beadle

Forsvarets forskningsinstitutt (FFI) 14. april 2021

Emneord
Sikkerhetspolitikk
Forsvarspolitikk
Prediksjon
Framtidsstudier

FFI-rapport
21/00735

Prosjektnummer
1553

Elektronisk ISBN
978-82-464-3347-9

Engelsk tittel
How to improve the accuracy of predictions in international politics? – a literature review

Godkjennere
Alf Christian Hennum, *forskningsleder*
Sigurd Glærum, *forskningssjef*

Dokumentet er elektronisk godkjent og har derfor ikke håndskreven signatur.

Sammendrag

Eksisterende forskning om hvor presist det er mulig å forutsi konkrete politiske hendelser, som utfallet av Brexit-avstemningen, antall nordkoreanske atomprøvesprengninger og hvor raskt Kinas økonomi vil vokse, baserer seg i hovedsak på to store, amerikanske forskningsprosjekter: *Expert Political Judgment* (EPJ) fra 2005 og *Good Judgment Project* (GJP) fra 2011–2015.

På den ene siden var funnene fra EPJ nedslående. Her ble treffsikkerheten til 284 eksperter målt på spørsmål som så 2, 5, 10 eller 20 år fremover. Ekspertene slet med å slå tilfeldig gjetning når tidsperspektivet nærmet seg 3–5 år. Det viste seg også at utdannings- og erfaringsnivå hadde lite å si for treffsikkerheten. Selv eksperter som predikerte innenfor sine egne områder, traff ofte dårligere enn andre eksperter som predikerte utenfor sitt.

På den annen side var resultatene fra GJP-prosjektet langt mer lovende. GJP var ett av lagene som deltok i en fireårig turnering sponset av amerikansk etterretning. For å treffe best mulig ble det forsket på ulike metoder for å aggregere prediksjoner fra tusenvis av deltagere. Det ble stilt flere hundre spørsmål med et tidsperspektiv på rundt 100 dager i snitt. Allerede etter to år traff GJP så godt at andre lag ble lagt ned. Funnene fra GJP viste at det er mulig å forutsi utfall på spørsmål av betydning for amerikansk etterretning. Vinneroppskriften var en kombinasjon av å rekruttere de riktige folkene og tiltak som forbedret treffsikkerheten.

Et gjennomgående funn i både EPJ og GJP var at noen personer i utgangspunktet er bedre til å predikere enn andre. Bedre treffsikkerhet hang sammen med høyere score på tester av kognitive evner, politisk kunnskapsnivå og fordomsfri tenkning. De aller beste hadde også et høyere ønske om å treffe best, interesse for mentalt krevende aktiviteter og en mer vitenskapelig tilnærming til det å vurdere fremtidige hendelser. Samtidig fant GJP at det var mulig å forbedre treffsikkerheten ytterligere gjennom tre tiltak: 1) opplæring i probabilistisk tenkning, som bruk av grunnfrekvens, 2) interaksjon mellom deltagerne, både i form av samarbeid i grupper og av konkurranse i prediksjonsmarkeder og 3) algoritmer som vektla prediksjonene til personer som har truffet godt tidligere og nylig som har oppdatert prediksjonene sine.

Funnene fra EPJ og GJP er imidlertid ikke nødvendigvis overførbare til en norsk forsvars- og sikkerhetspolitisk kontekst. Deltagerne i begge studier var stort sett amerikanske. Det er ikke gitt at de samme individuelle variasjonene vil gjelde for norske eksperter og deltagere. Det er heller ikke gitt at funnene vil være de samme om spørsmålene tar utgangspunkt i de viktigste aktørene for norsk sikkerhet. Selv om ekspertene i EPJ slet mot tilfeldig gjetning når tidsperspektivet nærmet seg 3–5 år, traff også de bedre jo kortere tidsperspektivet var. I GJP var tidsperspektivet på rundt 100 dager langt kortere og dermed enklere å treffe innenfor. Hensikten med FFIs prediksjonsturnering (2017–2020), som denne rapporten danner det teoretiske grunnlaget for, var derfor å etterprøve funnene fra EPJ og GJP med norske deltagere på spørsmål av betydning for norsk sikkerhet og med et tidsperspektiv på mellom 100 dager og 3–5 år.

Summary

Existing research on the accuracy of predictions in international politics, such as the outcome of the Brexit vote, the number of North-Korean nuclear weapons tests and the growth rate of the Chinese economy, is largely based on two research projects conducted in the US: *Expert Political Judgment* (EPJ) from 2005 and the *Good Judgment Project* (GJP) from 2011–2015.

On the one hand, the findings from EPJ were depressing. Here, the accuracy of 284 experts was measured on questions that looked 2, 5, 10 or 20 years ahead. The experts struggled to beat guessing when the time perspective approached 3–5 years. It was also found that levels of education or years of experience did not correlate with accuracy. Experts predicting inside their own domains of expertise were also often worse than those predicting outside theirs.

On the other hand, the results from GJP were far more encouraging. GJP was one of the teams participating in a four-year forecasting tournament sponsored by US intelligence. In order to achieve the highest possible accuracy, researchers experimented with various methods for aggregating predictions from thousands of participants. Hundreds of questions were posed, with an average time perspective of around 100 days. After only two years, GJP did so well that the other teams were dropped. The findings from GJP showed that it was possible to predict the outcome of questions of relevance to US intelligence. The winning recipe was a combination of recruiting the right people and taking measures that helped improve the overall accuracy.

A common finding in both EPJ and GJP was that there are systematic individual differences in accuracy. Better accuracy was associated with higher scores on tests of cognitive abilities, political knowledge and open-minded thinking. The best forecasters were also more motivated by the desire to win, had a higher need for cognition and a more probabilistic approach to future events. At the same time, GJP found that it was possible to improve accuracy through several measures: 1) training in probabilistic thinking, e.g. the use of base rates; 2) interaction between participants, both in the form of cooperation in groups and competition through prediction markets; and 3) algorithms that weighted the predictions made by participants who had previously been more accurate and who had recently updated their forecast.

However, these findings are not necessarily valid in a Norwegian defence and security policy context. Participants in both studies were largely US citizens. It is not given that the same individual variations exist among Norwegian experts and participants. Neither is it given that the results will hold on questions on the most important actors to Norwegian national security. Even though experts in EPJ struggled to beat guessing on questions that looked 3–5 years ahead, they were more accurate the shorter the time perspective. Thus, GJP's time perspective of 100 days was likely easier to forecast within. The purpose of FFI's forecasting tournament (2017–2020) was therefore to examine these findings with Norwegian participants on questions of relevance to Norway and with a time perspective between 100 days and 3–5 years.

Innhold

Forord

Denne rapporten er skrevet som en del av FFIs forskningsprosjekt «Globale trender og militære operasjoner III» (2019–2022). Hensikten med forskningsprosjektet er å styrke FFIs støtte til Forsvarets langtidsplanlegging med et langsiktig perspektiv på globale utviklingstrekk og å drive kompetansebygging på militære fremtidsstudier for Forsvaret.

Rapporten er et bidrag til begge målsettinger gjennom å beskrive den nyeste forskningen om hvor langt og presist det er mulig å forutsi konkrete politiske hendelser og utviklinger, og ikke minst hvordan det er mulig å forbedre treffsikkerheten i forbindelse med etterretningsvurderinger og utredninger til støtte for Forsvaret.

Deler av forskningen som beskrives her, er oppsummert i tidligere FFI-rapporter, men det har siden da blitt publisert flere nye studier. Det har derfor vært nødvendig med en oppdatert og mer detaljert gjennomgang av forskningen, som også danner det teoretiske utgangspunktet for hvilke hypoteser som har blitt testet i FFIs egen prediksjonsturnering (2017–2020).

Alexander W. Beadle
Kjeller, 14. april 2021

1 Innledning

De siste årene er det gjort store fremskritt innenfor forskningen på prediksjon av politiske hendelser og utviklinger, hvilke personer som er bedre til å predikere enn andre, hva som kjennetegner de aller beste og hvordan det er mulig å forbedre treffsikkerheten.

Denne forskningen kommer hovedsakelig fra to store, amerikanske forskningsprosjekter. Det første er *Expert Political Judgment* (EPJ), som målte treffsikkerheten til profesjonelle eksperter på geopolitiske spørsmål med flere års tidsperspektiv. Det andre prosjektet er *Good Judgment Project* (GJP). GJP var ett av lagene som deltok i en fireårig prediksjonsturnering arrangert av amerikansk etterretning, men som slo alle de andre med klar margin allerede etter to år. Her ble det samlet flere hundre tusen prediksjoner fra rundt 3000 deltagere, på rundt 500 spørsmål om et bredt spekter av politiske temaer, med noen måneders tidsperspektiv.

De til dels overraskende funnene fra disse forskningsprosjektene var bakgrunnen for at FFI arrangerte en egen prediksjonsturnering fra 2017 til 2020. Hensikten var å måle treffsikkerheten til det norske forsvars- og sikkerhetspolitiske miljøet og å etterprøve de to prosjektenes funn om hvem som treffer bedre enn andre på spørsmål av relevans for forsvarsplanlegging. Spørsmålene og resultatene fra FFIs turnering er beskrevet i egne rapporter,[1] men forskningen som oppsummeres her, danner det teoretiske grunnlaget for hypotesene som ble undersøkt.

Det er spesielt funnene fra GJP som er viet mest plass i denne rapporten. Mens hensikten i EPJ var begrenset til å måle hvor godt det er mulig å predikere, skulle GJP identifisere tiltak som bidrar til å *forbedre* treffsikkerheten. Det ble derfor gjennomført en rekke eksperimenter, som å sette deltagere sammen i grupper eller gi dem opplæring i teknikker for å unngå vanlige feilslutninger. Flere av disse tiltakene er relevante i norsk sammenheng, og det er derfor hensiktsmessig å beskrive disse funnene, selv om de går utover hypotesene som ble testet i FFIs turnering.

Kapittel 2 innleder rapporten ved å beskrive utfordringer ved dagens bruk av rent verbale sannsynlighetsvurderingene i forsvars- og sikkerhetspolitisk sammenheng. Kapittel 3 og 4 oppsummer funnene fra hvert av de to amerikanske forskningsprosjektene. Her viser det seg at formelle kvalifikasjoner, som ofte brukes til å selektere fagfolk til bruk i utredninger og i media, har lite å si for prediksjonsevnen. I stedet varierer treffsikkerheten med andre individuelle egenskaper, som ulike mål på kognitive evner, tenkemåter og adferd i prediksjonssammenheng.

Kapittel 5 beskriver pågående forskningsprosjekter. Kapittel 6 avslutter med å oppsummere implikasjonene av eksisterende forskning for mulighetene for å forbedre treffsikkerheten i norsk forsvarssammenheng, inkludert hvilke forbehold som må tas og undersøkes nærmere.

[1] Beadle, A. W. (2021), 'FFIs prediksjonsturnering – spørsmålskatalog', *FFI-rapport 21/00736* (Kjeller: Forsvarets forskningsinstitutt); Beadle, A. W. (2021), 'FFIs prediksjonsturnering – datagrunnlag og foreløpige resultater', *FFI-rapport 21/00737* (Kjeller: Forsvarets forskningsinstitutt).

2 Bakgrunn

Å predikere, eller forutsi, handler om å beskrive en fremtidig utvikling eller hendelse. Ofte inkluderer prediksjoner også en vurdering av hvor sannsynlig noe anses å være. For eksempel vurderte Etterretningstjenesten det i 2020 som «sannsynlig» at Russland ville øke sitt diplomatiske engasjement i Libya, «trolig» at Iran ville forhandle med USA hvis sanksjonene ble lettet og «mulig» med en nordkoreansk prøvesprengning.[2] I løpet av året som gikk, var det først og fremst økt militær støtte til general Khalifa Haftar som preget analysene av Russlands engasjement i Libya. Det ble heller ikke gjennomført lettelser av sanksjonene mot Iran eller en nordkoreansk prøvesprengning.

I forsvarsplanlegging er det ikke vanlig med så eksplisitte prediksjoner, men det gjøres likevel sannsynlighetsvurderinger. På den ene siden er FFIs scenarioporteføljе ment å fange et spekter av mulige trusler, nettopp fordi det er umulig å forutsi nøyaktig hva slags angrep Norge kan bli utsatt for. På den annen side baserer *utvalget* av scenarioer seg på vurderinger av hva som anses som «mulig».[3] For eksempel legges det til grunn at Russland har evnen og viljen til å gjennomføre et begrenset angrep, men ikke til å invadere hele Norge, slik en trodde om Sovjetunionen.[4]

Felles for de fleste prediksjoner som gjøres i forsvars- og sikkerhetspolitisk sammenheng, er at de sjelden tallfestes. I stedet er det vanligste verbale, kvalitative formuleringer, som i eksemplene over. Dette er naturlig fordi politiske hendelser og utviklinger er vanskelige å beregne statistisk, både fordi de er komplekse, sosiale fenomener og fordi de samme typene hendelsene sjeldent skjer mer enn én gang.

Problemet er at verbale sannsynlighetsvurderinger kan gjøre det vanskeligere for beslutningstagere å vite hva som egentlig menes. Studier har for eksempel vist at ordet «mulig» oppfattes som mindre enn 10 % sannsynlig av noen, og som mer enn 50 % sannsynlig av andre.[5] Dette medfører en reell fare for at viktige beslutninger fattes på feil grunnlag.

Ett av de mest kjente eksemplene er USAs mislykkede invasjon av Cuba i 1961, der president John F. Kennedy fikk beskjed om at planen hadde en «*fair chance*» for å lykkes. Mannen bak ordene «*fair chance*» har senere uttalt at han med dette mente at det var omtrent 1/3 sjanse for at planen ville lykkes. Denne tallfestingen fikk aldri Kennedy. I stedet tolket han formuleringen som at det var en høyere sannsynlighet for at planen ville lykkes enn ikke, og iverksatte den.

[2] For eksempel gir varianter av «sannsynlig» 110 treff i Etterretningstjenesten åpne trusselvurdering, *Fokus 2020*.
[3] Johansen, I. (2006), 'Scenarioklasser i Forsvarsstudie 2007: En morfologisk analyse av sikkerhetspolitiske utfordringer mot Norge', *FFI-rapport 2006/02664* (Kjeller: Forsvarets forskningsinstitutt).
[4] Åtland, K., Beadle, A. W., Diesen, S., Glærum, S., Mørkved, T., Nyhamar, T. og Stenersen, A. (2018), 'Gjennomgang av FFIs scenariogrunnlag for Forsvarets langtidsplanlegging, 2018', *FFI-rapport 18/00669* (Kjeller: FFI). (BEGRENSET).
[5] For en studie av de forskjellige prosentvise verdiene som akademikere tillegger verbale sannsynlighetsformuleringer, se Mosteller, F. og Youtz, C. (1990), 'Quantifying Probabilstic Expressions', *Statistical Science*, 5:1, ss. 2–12. Eksempelet her er basert på forskjellen mellom 25. og 75. kvartil for ordet «possible» («mulig») i tabell 2, s. 6.

Et annet problem med rent verbale sannsynlighetsvurderinger er at vi risikerer å fortsette å predikere galt, uten at vi selv er klar over det. Vi har alle en tendens til å underspille egne feilprediksjoner og å finne forklaringer på hvorfor vi hadde rett, uansett hva vi trodde før. Uten tallfestede sannsynligheter i forkant er det derfor alltid mulig å strekke formuleringer som «mulig» til riktig side av svaret i etterkant. Konsekvensen er at vi alltid kan få «rett» og bare fortsetter å legge gale antagelser til grunn i analysene våre.

Mangelen på tallfestede prediksjoner av utviklingen i internasjonal politikk har også gjort det vanskelig å forske på hvor godt de egentlig treffer. Det var nettopp dette gapet i eksisterende forskning som de to prosjektene som beskrives i denne rapporten ønsket å adressere.

3 *Expert Political Judgment* (EPJ)

Fra 1980-tallet og frem til 2003 samlet professor Philip E. Tetlock inn prediksjoner fra 284 politiske eksperter. Her ble ekspertene bedt om å anslå sannsynligheten (i antall prosent) til politiske, økonomiske og sikkerhetspolitiske utviklinger, både innenfor og utenfor sine egne fagområder. Dette arbeidet ble publisert i boken *Expert Political Judgment* (EPJ), som kom ut i 2005.[6]

Alle deltagerne i EPJ var «profesjonelle eksperter» som arbeidet med trender av betydning for stater, regioner eller verden generelt.[7] Alle deltok anonymt. Flesteparten var menn (76 %). Gjennomsnittsalderen var 43 år. De hadde i snitt 12 års relevant arbeidserfaring, 52 % hadde doktorgrad og 96 % hadde utdanning på mastergradsnivå. Faglig kom de fleste fra områdestudier (41 %), internasjonal relasjoner (24 %), økonomi (12 %) eller nasjonal sikkerhet og rustningskontroll (11 %). De arbeidet hovedsakelig i akademia (41 %), staten (26 %) eller tenketanker og stiftelser (17 %). Rundt 61 % hadde blitt intervjuet av minst ett stort medium og 21 % hadde blitt intervjuet minst 10 ganger. Rundt 80 % hadde bistått myndigheter, private, organisasjoner eller tenketanker med analyser av internasjonal politikk eller økonomi.

EPJ var først og fremst en geopolitisk prediksjonsstudie.[8] Ekspertene kunne få spørsmål om utviklingen i 60 land fordelt på 9 regioner (Sovjetunionen, Europa, Nord-Amerika, Mellom og Latin-Amerika, den arabiske verden, Afrika sør for Sahara, Kina, Nordøst-Asia og Sørøst-Asia). Ekspertene ble bedt om å gjøre én kortsiktig og én langsiktig prediksjon om utviklingen i 4 ulike land, hvorav 2 lå innenfor og 2 utenfor sitt eget kompetanseområde. For hvert land ble de

[6] Tetlock, P. (2005), *Expert Political Judgment: How Good Is It? How Can We Know?* (Princeton: Princeton University Press). Denne oppsummering er basert på kapittel 2 og 3 som beskriver ekspertenes geopolitiske prediksjoner.
[7] For definisjonen av «ekspert» og mer informasjon om dem, se Tetlock (2005), *Expert Political Judgment*, ss. 239ff.
[8] For mer informasjon om spørsmålene og prediksjonene, se Tetlock (2005), *Expert Political Judgment*, ss. 239–252.

bedt om å oppgi ett sannsynlighetsestimat for tre forskjellige utfall på 17 ulike områder i gjennomsnitt. Dette utgjorde rundt 140 spørsmål med tre utfall hver per ekspert.[9] Til sammen ble det samlet inn 82 361 sannsynlighetsvurderinger fra de 284 ekspertene.[10]

Spørsmålene kunne dreie seg om fire temaer: 1) politisk styring og stabilitet, som valgresultater og kupp, 2) innenrikspolitisk og økonomisk utvikling, som bruttonasjonalprodukt og rentenivåer, 3) forsvars- og sikkerhetspolitikk, som deltagelse i militære operasjoner og allianser og 4) forskjellige casestudier, som spredningen av masseødeleggelsesvåpen og maktskifter i tidligere kommunistland. En ekspert kunne for eksempel bli bedt om å vurdere sannsynlighetene for at det regjerende politiske partiet i et land innenfor hans kompetanseområde ville få større, mindre eller omtrent lik oppslutning, både ved det neste valget (kort sikt) og et senere valg (lang sikt). De fleste spørsmålene ba ekspertene predikere 2, 5, 10 eller 20 år frem. Da studien ble publisert, var det imidlertid bare noen av spørsmålene som så 10 år eller lenger fremover, som var avgjort.

Resultatene var likevel nedslående: Ekspertene klarte bare så vidt å slå tilfeldig gjetning, der en bare hadde fordelt sannsynligheten helt likt på alle utfall på alle spørsmål (f.eks. 50/50 på et spørsmål med to utfall). De beste ekspertene traff på nesten 60 %, som var bedre enn gjetning, men ikke mye.[11] De dårligste ekspertene slet imidlertid med å slå tilfeldig gjetning. Dette gav opphav til utsagnet som prosjektet ble mest kjent for: at eksperter var like dårlige til å predikere som en pilkastende ape med bind for øynene, der det er helt tilfeldig hvor godt man treffer.[12]

Tetlock har siden studien ble publisert ønsket å nyansere dette inntrykket.[13] For det første spilte spørsmålenes relativt lange tidsperspektiv en rolle. Treffsikkerheten nærmet seg tilfeldig gjetning på spørsmål som så rundt fem år fremover, men ekspertene traff bedre jo kortere tidsperspektivet var.[14] For det andre var ikke alle ekspertene like dårlige. Tvert imot fant Tetlock at det var mulig å skille mellom to stereotyper eksperter – pinnsvin og rever – basert på *hvordan* de tenkte:[15]

- *Pinnsvinene* var kjennetegnet av at de kunne ett eller to store emner eller teorier, som globalisering, maktbalanseprinsippet eller sivilisasjonskonflikt, som de appliserte på alle spørsmål (deduktiv resonnering). De plasserte komplekse problemer inn i årsak-virkningsforhold som de kjente fra før, mens det som ikke passet inn ble behandlet som irrelevant. Pinnsvinene var svært selvsikre i sine prediksjoner, og hadde lettere for å avvise motsigende synspunkter. De brukte gjerne ord som «dessuten», «og så videre» og «i tillegg til» for å trekke inn ytterligere argumenter for hvorfor de hadde rett, og

[9] Antall spørsmål er ikke oppgitt i boken, men ble av opplyst av Tetlock selv gjennom korrespondanse 14. des. 2020.
[10] Tetlock (2005), *Expert Political Judgment*, s. 246.
[11] 'A talk with Philip Tetlock', *Boston Globe*, 5. okt. 2008, sitert i 'Research That Makes You Go Hmmm on…Forecasts and Predictions', *The Clemmer Group*, 12. jan. 2016.
[12] På engelsk: *dart-throwing chimpanzee*. For en diskusjon av metaforen, se forordet og s. 68 i Tetlock, P. og Gardner, D. (2015), *Superforecasting: The Art and Science of Prediction* (London: Random House Books).
[13] For en oppdatert oppsummering av funnene fra EPJ, se forordet i den nye utgaven av boken, Tetlock, P. E. (2017), *Expert Political Judgment: How Good Is It? How Can We Know?* (New Jersey: Princeton University Press).
[14] Tetlock og Gardner (2015), *Superforecasting*, s. 5 og s. 244.
[15] For mer om de to stereotypene eksperter, se kapittel 3–6 i Tetlock (2005), *Expert Political Judgment*, og Teltock og Gardner (2015), *Superforecasting*, ss. 68–73.

skydde ikke ord som «umulig» eller «sikkert» i sine vurderinger av fremtiden. Gale prediksjoner ble bortforklart ved at de «bommet litt på tidspunktet», var «nesten riktige» eller at de ble avsporet av «uforutsigbare» hendelser.

– *Revene* var derimot kjennetegnet av at de kunne mange forskjellige, men ikke så store ting. De var skeptiske til store idéer om hvordan verden henger sammen og hvilke lover som egentlig gjaldt. I stedet anvendte de forskjellige analytiske tilnærminger avhengig av problemet som skulle løses (induktiv resonnering). De samlet mer informasjon fra mange kilder før de bestemte seg. I språket sitt brukte de oftere ord som «men», «imidlertid», «selv om» og «på den annen side». De snakket også om muligheter og sannsynligheter, ikke sikkerheter – og hadde lettere for å innrømme feil.

Av disse to stereotypene var reveekspertene mye bedre til å predikere enn pinnsvinekspertene.[16] Pinnsvinene gjorde det faktisk ofte dårligere enn apen. Revene slo apen, men klarte likevel bare så vidt å slå enkle algoritmer som predikerte «ingen endring» eller «dagens endringstempo».

Et overraskende funn var at ekspertenes utdannings- og erfaringsnivå hadde svært lite å si for variasjoner i treffsikkerheten. Det spilte ingen rolle om ekspertene hadde doktorgrad, politisk erfaring eller tilgang på gradert informasjon, hvorvidt de var økonomer, statsvitere, journalister eller historikere, eller hvor mange års erfaring de hadde innenfor deres egen profesjon.[17] Dette er spesielt relevant for beslutningstagere fordi slike formelle kriterier ofte brukes til å selektere eksperter til utredninger av fremtidige trusler og behov. Selv ikke ekspertene som predikerte innenfor sine egne områder, traff bedre enn andre eksperter som predikerte utenfor.

En forklaring på fraværet av sammenheng mellom formell kompetanse og prediksjonsevne er at ekspertise og erfaring betyr mindre når usikkerheten uansett er stor, slik det ofte er med politikk. Her skiller politikk seg fra andre fagfelt. Meteorologer og profesjonelle sjakkspillere er åpenbart bedre til å vurdere sannsynligheten av fremtidige utfall enn amatører.[18] Erfarne brannmenn og jordmødre klarer også å vurdere situasjoner raskere enn ferske nybegynnere. Felles for disse yrkene er at de befinner seg i «lærevennlige» verdener. Her får en raskt får vite hvor godt en traff, som igjen gjør det mulig å forbedre senere prediksjoner. Innenfor politikk er dette langt vanskeligere. Politiske eksperter predikerer ofte forhold som er vanskelige å kvantifisere, de må som regel vente lenge før de får vite hva utfallet ble, og selv da er svarene fortsatt åpne for ulike tolkninger. I slike omgivelser preget av stor usikkerhet, er mennesker mer utsatt for psykologiske mekanismer som leder til gale svar, som at vi hopper til konklusjoner for raskt, endrer mening for sent, overreagerer på små, nye detaljer og tror på noe bare fordi andre gjør det.[19]

Revene skilte seg fra pinnsvinene ved at de i mindre grad gikk i slike psykologiske fallgruver. Revene hadde en høyere toleranse for usikkerhet. De var mindre farget av sine egne, ideologiske oppfatninger når de skulle predikere. Revene erkjente også at usikkerheten økte jo lenger frem en skal predikere, og var mer åpen for at uforutsigbare hendelser kunne dukke opp og overraske

[16] Tetlock og Gardner (2015), *Superforecasting*, s. 68ff.
[17] Tetlock (2005), *Expert Political Judgment*, s. 68.
[18] Se forordet i Tetlock (2017), *Expert Political Judgment*, for referanser på dette.
[19] For en oppsummering av forskningen, se Kahneman, D. (2013), *Tenke, fort og langsomt* (Oslo: Pax Forlag).

selv de beste deltagerne. Pinnsvinene ble derimot bare mer tilbøyelig til å applisere sine overordnede, deduktive teorier jo lenger frem de skulle predikere og jo større usikkerheten ble.

Den eneste bakgrunnsvariabelen som hang sammen med treffsikkerhet var ekspertenes berømmelse (målt som antall Google-treff). Sammenhengen var imidlertid omvendt: Jo *mer kjent* eksperten var, jo *dårligere* var treffsikkerheten. En mulig forklaring var ifølge Tetlock at mediene foretrekker eksperter som er bastante og selvsikre (som pinnsvinene), og at disse ekspertene dermed blir brukt oftere, selv om de er dårligere til å predikere. Eksperter blir også sjeldent vurdert ut fra tidligere treffsikkerhet, fordi denne evnen sjelden måles. Det var dette Tetlock ønsket å gjøre noe med i EPJ-studien, men resultatene var ikke oppløftende.[20]

4 *Good Judgment Project* (GJP)

Fra 2011 til 2015 ble det gjennomført en omfattende, fireårig prediksjonsturnering i USA. Turneringen ble arrangert av den føderale etaten *Intelligence Advanced Research Projects Activity* (IARPA), som sponser forskningsprosjekter som bidrar til å løse spesielt vanskelige utfordringer for amerikansk etterretning. Hensikten var å identifisere metoder som kunne øke treffsikkerheten i etterretningsanalyser. Fem lag fra akademia og industri konkurrerte om hvem som var best til å predikere svarene på rundt 500 spørsmål om internasjonal politikk, som for eksempel: Vil Nord-Korea detonere et atomvåpen de neste tre månedene? Hvor mange flyktninger vil flykte fra Syria det neste året? Hvor raskt vil Kinas økonomi vokse det neste kvartalet?

De to første årene ble alle spørsmålene utarbeidet av IARPA. Spørsmålene skulle være representative for det etterretningstjenestene typisk måtte svare på, men unngikk amerikansk innenrikspolitikk.[21] Spørsmålene måtte også kunne avgjøres innen «rimelig tid» (som regel under ett år) og være «tilstrekkelig vanskelige».[22] Spørsmål ble ansett som for enkle å predikere hvis sannsynligheten for at hendelsen ville skje ble vurdert som mindre enn 10 % sannsynlig eller mer 90 % sannsynlig ved tidspunktet spørsmålet ble stilt.[23] Målet var å spørre om hendelser med en sannsynlighet rundt midten av skalaen (50 %).

[20] For en oppsummering av kritiske bemerkninger til EPJ og Tetlock's svar på disse, se Tetlock, P. E. (2010), 'Second Thoughts about Expert Political Judgment: Reply to the Symposium', *Critical Review*, 22: 4, ss. 467–488.
[21] Mellers, B., Tetlock, P. og Arkes, H. R. (2019), 'Forecasting tournaments, epistemic humility and attitude depolarization', *Cognition*, 188, ss. 19–26, s. 22.
[22] Moore, D. A., Swift, S. A., Minster, A., Mellers, B., Ungar, L., Tetlock, P., Yang, H. H. J. og Tenneye, E. R. (2017), 'Confidence Calibration in a Multiyear Geopolitical Forecasting Competition', *Management Science*, 63:11, ss. 3552–3565, ss. 3555.
[23] Atanasov, P., Rescober, P., Stone, E., Swift, S. A., Servan-Schreiber, E., Tetlock, P., Ungar, L., og Mellers, B. (2017), 'Distilling the Wisdom of Crowds: Prediction Markets vs. Prediction Polls', *Management Science*, 63:3, ss. 587–900, s. 592.

Lagene måtte selv rekruttere deltagere og bestemme hvordan de skulle samle inn prediksjoner. Nye spørsmål ble publisert i grupper på ca. fire–fem per uke. Hver dag måtte lagene rapportere en aggregert prediksjon til IARPA, som beregnet treffsikkerheten ved hjelp av Brier-score.[24] Ved lanseringen var IARPAs mål at lagene skulle slå et uvektet snitt av alle prediksjoner med 20 % det første året, 30 % det andre året, 40 % det tredje året og 50 % det fjerde året.

Ett av lagene som deltok var *Good Judgment Project* (GJP), som ble etablert av tidligere nevnte Tetlock og professor Barbara A. Mellers.[25] Deres tilnærming var å lage en egen turnering innad i prosjektet for å gjennomføre eksperimenter som testet ulike tiltak for å øke deltagernes og dermed den aggregerte treffsikkerheten som GJP leverte til IARPA-turneringen. GJPs interne turnering ble gjennomført online, der deltagerne måtte oppgi hvor sannsynlig (i prosent) de trodde ulike utfall var. Deltagerne kunne oppdatere sine prediksjoner helt frem til spørsmålet ble stengt. De kunne også velge hvilke spørsmål de skulle svare på, men ble oppfordret til å svare på så mange spørsmål som mulig. De konkurrerte med hverandre, enten alene eller på ulike lag.

Fire av de fem lagene i IARPA-turneringen hadde imidlertid vanskeligheter med å rekruttere og beholde nok deltagere.[26] GJP skilte seg ut ved å rekruttere tusenvis av deltagere. Turneringen var i utgangspunktet åpen for alle, men deltagerne måtte ha utdanning på minst bachelorgradsnivå og gjennomføre psykologiske og politiske kunnskapstester som tok rundt to timer til sammen. Av deltagerne var 83 % var menn, 78 % amerikanske statsborgere og medianalderen var 35 år.[27] Av deltagerne hadde 63 % utdanning på minst mastergradsnivå. Deltagerne var altså «eksepsjonelt godt utdannet, motiverte og informerte»,[28] men samtidig ikke rekruttert ut fra ekspertise innenfor temaene de skulle predikere, slik som i EPJ.

Deltagerne fikk to typer belønning.[29] Den første var status, basert på treffsikkerhet. Navnene på de 10 % beste ble offentliggjort på ledertavler innenfor hver eksperimentgruppe og innad på eventuelle lag. I tillegg fikk deltagerne se lagets plassering i forhold til andre. Den andre belønningen var betaling for deltagelse. Deltagere som predikerte minst 25 ganger per år fikk et gavekort ($150 det første året av turneringen, $250 det andre og tredje året). I tillegg fikk deltagere fra det første året som også deltok i det andre eller tredje året, en bonus på $100.

Til forskjell fra EPJ var resultatene fra GJP svært oppløftende. Allerede etter to år traff de aggregerte prediksjonene fra GJP så godt at de fire andre lagene i IARPA-turneringen ble lagt

[24] Brier-systemet er et av de vanligste målene på treffsikkerheten til probabilistiske prediksjoner. Her måles ikke prediksjonen ut fra *om* den treffer, men hvor *sannsynlig* (i prosent) det riktige utfallet anslås for å være. Skalaen går fra 0 til 2, der lavere score betyr høyere treffsikkerhet. Du får en Brier-score på 0 hvis du predikerer «helt riktig», det vil si at du hevder at en hendelse er 100 % sannsynlig, og den faktisk skjer. Du får en Brier-score på 2 hvis du predikerer «helt feil», det vil si at du hevder en hendelse er 100 % sannsynlig, men den *ikke* skjer. Se Brier, G. W. (1950), 'Verification of Forecasts Expressed in Terms of Probability', *Monthly Weather Review*, 78:1.

[25] For mer informasjon om GJP, se Tetlock og Gardner (2015), *Superforecasting*, ss. 16–20 og ss. 87–96. For et intervju med Tetlock, se 'How to Be Less Terrible at Predicting the Future', *Freakonomics*, 14. jan. 2016. Ifølge Tetlock var arbeidsdelingen mellom Mellers og ham at Mellers gjør den dype forskningen, mens tar seg av kommunikasjonsarbeidet. Mellers er også førsteforfatter på mange av artiklene basert på resultatene fra GJP.

[26] 'The Aggregative Contingent Estimation Program', *CitizenScience.gov*.

[27] Moore et al. (2017), 'Confidence Calibration in a Multiyear Geopolitical Forecasting Competition', s. 3555.

[28] Ibid. s. 3563.

[29] Ibid. s. 3555–3557.

ned.[30] Deretter overtok GJP ansvaret for spørsmålsgenereringen og kunne rekruttere deltagere fra de andre lagene, som gjorde at det samlede antallet deltagere økte betydelig.

De to siste årene av prosjektet ble brukt til å optimalisere metodene som hadde vist seg å være lovende for å oppnå høyest mulig aggregert treffsikkerhet. Oppsummert bestod vinneroppskriften til GJP av fire tiltak, som hver for seg økte den samlede prediksjonsevnen:[31]

1) *Rekruttering av de beste deltagerne*. EPJ hadde allerede funnet at personer med reveaktige egenskaper var bedre til å predikere. På grunn av denne forskningen hadde Tetlock også fordelen av å være mer kjent, som ifølge ham selv gjorde at GJP tiltrakk seg bedre deltagere enn de andre lagene. Rekrutteringen av de riktige folkene ble anslått å ha økt treffsikkerheten til GJP med rundt 10–15 % sammenlignet med andre lag.

2) *Opplæring i probabilistisk tenkning*.[32] Basert på tidligere forskning fra kognitiv psykologi om hvordan mennesker er spesielt utsatt for å tankefeil når de skal gjøre vurderinger preget av stor usikkerhet, utviklet GJP undervisningsmoduler som skulle hjelpe deltagerne til å unngå vanlige fallgruver, som ønsketenkning, bekreftelsestendensen og etterpåklokskap. Modulene ble videreutviklet hvert år, men la særlig vekt på hvordan tenke probabilistisk (ved hjelp av sannsynligheter), for eksempel bruk av grunnfrekvens (hvor hyppig et fenomen er), referanseklasser (undersøke tidligere utfall i lignende situasjoner),[33] statistiske modeller for oppdatering av prediksjoner i lys av ny informasjon (f.eks. Bayes teorem) og å basere seg på gjennomsnittet av flere, uavhengige vurderinger. Til tross for at ingen av modulene varte i mer enn én time, anslås denne opplæringen å ha økt deltagernes treffsikkerhet med 6–11 % over kontrollgruppen.[34]

3) *Interaksjon mellom deltagerne*.[35] GJP eksperimenterte med to former for interaksjon mellom deltagerne, basert på en antagelse om at deltagere som delte informasjon med hverandre ville treffe bedre enn deltagere som predikerte alene. Bakgrunnen er at forskjellige personer kan besitte ulike deler av informasjon som kan være relevante, og at

[30] Da IARPA-turneringen ble lansert, var målet å slå et uvektet snitt av alle prediksjoner med 20 % det første året, 30 % det andre året, 40 % det tredje året og 50 % det fjerde året. GJPs beste forecastere og beste algoritmer slo målet om 50 % allerede etter det første året, og de fortsatte å gjøre det de neste tre årene. GJP var det eneste laget som konsekvent slo IARPAs mål for de to første årene. Se kursserien 'Edge Master Class 2015: A Short Course in Superforecasting', *Edge*, 17. aug. 2015–21. sept. 2015 for mer informasjon om gjennomføringen og resultatene.

[31] For en oppsummering av tiltakene, se Tetlock, P., Mellers, B., Rohrbaugh, N. og Chen, E. (2014), 'Forecasting Tournaments: Tools for Increasing Transparency and Improving the Quality of Debate', *Current Directions in Psychological Science*, 23:4, ss. 290–295; Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., Scott, S. E., Moore, D., Atanasov, P., Swift, S., A., Murray, T., Stone, E. og Tetlock, P. E. (2014), 'Psychological strategies for winning a geopolitical forecasting tournament', *Psychological Science*, 25:4, 1106–1115. Det er noen forskjeller i de prosentvise anslagene på hvor mye hvert tiltak bidro til å øke treffsikkerheten som er oppgitt i Tetlock et al. (2014), Forecasting Tournaments' og på Tetlocks plansjer fra kursserien 'Edge Master Class 2015'. Hvis anslagene varierer er de to estimatene adskilt med bindestrek her.

[32] Chang, W., Chen, E., Mellers, B. og Tetlock, P. (2016), 'Developing expert political judgment: The impact of training and practice on judgmental accuracy in geopolitical forecasting tournaments', *Judgment and Decision Making*, 11:5, ss. 509–526.

[33] Kahneman og Tversky (1977), 'Intuitive prediction: Biases and corrective procedures', *Technical Report PTR-1042-77-6* (Virginia: DARPA).

[34] Se tabell 4, s. 515, i Chang et al. (2016), 'Developing expert political judgment'.

[35] Atanasov et al. (2017), 'Distilling the Wisdom of Crowds: Prediction Markets vs. Prediction Polls'.

flere studier har vist at snittet av mange prediksjoner («*wisdom of the crowd*») ofte er mer treffsikkert enn de fleste enkeltpersoners.[36] Dette skyldes blant annet at betydningen av enkeltpersoners potensielt svært gale svar blir utlignet av det store antallet prediksjoner.

Den første formen for interaksjon var gruppearbeid, der en del av deltagerne ble tilfeldig fordelt på grupper med opptil 15–25 personer.[37] Innad i gruppen kunne deltagerne dele artikler, utveksle argumenter og motivere hverandre. Gruppene ble designet slik at fordelene ved gruppearbeid, som større tilfang av mangfoldig kunnskap, skulle veie opp for ulempene, som faren for gruppetenkning.[38] Alle gruppemedlemmer ble derfor oppfordret til å begrunne sine prediksjoner og diskutere dem med de andre på laget. Deltagerne predikerte hver for seg, men gruppens aggregerte prediksjon baserte seg på medianen av alle medlemmenes. Hovedfunnet var at gruppearbeid økte treffsikkerheten. Resultatene viste at deltagere som arbeidet i grupper, traff bedre enn dem som predikerte alene. Grupper som samarbeidet mer og tenkte mer probabilistisk var også bedre. Konklusjonen ble at å sette personer sammen i grupper, som samarbeidet online, anonymt og med treffsikkerhet som eneste statusmarkør, bidro til å øke treffsikkerheten. En del av forklaringen kan være at deltagerne aldri møttes ansikt til ansikt, som kan ha motvirket faren for gruppetenkning.

Den andre formen for interaksjon var prediksjonsmarkeder, der deltagerne konkurrerte ved å kjøpe og selge aksjer som på vanlige børser. Her ble prisen brukt som markedets aggregerte prediksjon av fremtidige utfall, og deltagerne ble målt ut fra hvor mye de «tjente» på å predikere riktig. Prediksjonsmarkedene traff bedre enn et vanlig snitt av prediksjonene til deltagere som arbeidet i grupper. Når prediksjonene fra grupper ble aggregert på måter som la mer vekt på prediksjonene til deltagere som hadde truffet best tidligere, slo imidlertid gruppene prediksjonsmarkedet med god margin.

Forskjellen på treffsikkerheten til grupper og prediksjonsmarkeder avhenger derfor av hvordan de brukes, men både den samarbeidende og konkurrerende formen for interaksjon slo prediksjonene til deltagere som arbeidet alene. Ifølge Tetlock bidro gruppearbeid og prediksjonsmarkeder med å øke treffsikkerheten omtrent like mye, med rundt 10–20 % sammenlignet med deltagere som predikerte alene.

4) *Vekting av de beste deltagernes prediksjoner*. På samme måte som enkeltpersoners potensielt katastrofale vurderinger blir utlignet ved å aggregere mange prediksjoner, vil et uvektet snitt også utligne betydningen av prediksjonene til de aller beste deltagerne. GJP utforsket derfor algoritmer for å aggregere prediksjoner på måter som la større vekt

[36] Surowiecki, J. (2005), *The Wisdom of Crowds* (NY: Anchor Books).
[37] Det første året var gruppene på opptil 25 personer, mens det andre året var de på opptil 15 personer. Se Horowitz, M., Stewart, B. M., Tingley, D., Bishop, M., Samotin, L. R., Roberts, M., Chang, W., Mellers, B. og Tetlock, P. (2019), 'What Makes Foreign Policy Teams Tick: Explaining Variation in Group Performance at Geopolitical Forecasting', *The Journal of Politics*, 81:4, ss. 1388–1404.
[38] For en gjennomgang av potensielle ulemper ved gruppetenkning, se Mellers et al. (2014), 'Psychological strategies for winning a geopolitical forecasting tournament', s. 1107.

på svarene til de beste deltagerne. Dette ble f.eks. gjort ved å vektlegge prediksjonene til deltagere som hadde truffet bedre før og som hadde oppdatert prediksjonene sine nylig, fordi disse var antagelig basert på et mer oppdatert informasjonsgrunnlag.

Aggregeringsmetoden som traff best, baserte seg på et mål av hver enkelt deltagers bidrag til den samlede treffsikkerheten.[39] Her ble deltagernes bidrag vurdert som høyt, hvis fraværet av prediksjonene deres medførte et stort fall i gruppens treffsikkerhet. Dette målet ble oppdatert underveis i turneringen, basert på endringer i deltagernes bidrag. Dette representerte en potensielt kostnadsbesparende metode for prediksjon fordi den kan identifisere deltagere som ikke trenger å delta, ettersom prediksjonene deres bidrar lite til den samlede treffsikkerheten. Denne metoden viste seg også å være robust mot «sabotasje» fra deltagere som bevisst predikerte feil.

Basert på en vekting av de beste deltagernes prediksjoner «ekstremiserte» GJP de aggregerte prediksjonene ved å skyve sannsynlighetsestimatene nærmere det ene eller andre utfallet (0 % eller 100 %) enn de ville vært med et uvektet snitt.[40] Hvor mye prediksjonene burde ekstremiseres var avhengig av hvor mye av deltagernes informasjonsgrunnlag som overlappet. Hvis alle personer har samme informasjon, vil det i teorien ikke være behov for å ekstremisere den aggregerte prediksjonen. Hvis alle personer besitter ulik informasjon som peker i samme retning, har ekstremisering mye for seg. Prediksjoner fra personer som arbeider tett sammen har derfor lite å tjene på ekstremisering fordi informasjonsmengden deres overlapper mye, mens prediksjoner fra personer som har ulik informasjon og baserer seg på forskjellige kilder, burde ekstremiseres mer. GJP-forskerne har utviklet modeller for hvordan dette kan gjøres.[41]

Ifølge Tetlock bidro vektleggingen av de beste deltagernes prediksjoner og ekstremisering av de aggregerte sannsynlighetsvurderingene til å øke treffsikkerheten med 15–35 % sammenlignet med et vanlig, uvektet snitt av alle prediksjoner. Den beste aggregeringsalgoritmen til GJP havnet på riktig side av 50/50 på 86 % av alle daglige prediksjoner, som var langt bedre enn tilfeldig gjetning, og den slo snittet til vanlige deltagere som arbeidet alene uten trening med 60 % og andre lag med 40 %.[42]

[39] Chen, E., Budescu, D., Lakshmikanth, S., Mellers, B. og Tetlock, P. (2016), 'Validating the Contribution-Weighted Model: Robustness and Cost-Benefit Analyses', *Decision Analysis*, 13:2, ss. 128–152.
[40] For en teoretisk begrunnelse og empiriske beviser for å transformere aggregerte sannsynlighetsvurderinger mot det ekstreme, se Baron, J., Mellers, B. A., Tetlock, P. E., Stone, E. og Ungar, L. H. (2014), 'Two Reasons to Make Aggregated Probability Forecasts More Extreme', *Decison Analysis*, 11:2, ss. 133–145. For en nærmere beskrivelse av teknikker for aggregering av prediksjoner som ble utviklet i forbindelse med GJP, se Atanasov, P., Rescober, P., Stone, E., Servan-Schreiber, E., Mellers, B., Tetlock, P. og Ungar, L. (2013), 'The Marketcast Method for Aggregating Prediction Market Forecasts', i Greenberg, A. M., Kennedy, W. G., og Bos, N. D., red., *Social Computing, Behavioral-Cultural Modeling and Prediction* (SBP 2013); Satopää, V. A., Jensen, S. T., Mellers, B. A., Tetlock, P. E. og Ungar, L. H. (2014a), 'Probability aggregation in time-series: Dynamic hierarchical modeling of sparse expert beliefs', *The Annals of Applied Statistics*, 8:2, ss. 1256–1280; Satopää, V. A., Baron, J., Foster, D. P., Mellers, B. A., Tetlock, P. E. og Ungar, L. H. (2014b), 'Combining multiple probability predictions using a simple logit model', *International Journal of Forecasting*, 30:2, 344–356; Atanasov et al. (2016), 'Distilling the Wisdom of Crowds'.
[41] Satopää, V., Pemantle, R. og Ungar, L. (2015), 'Modeling Probability Forecasts via Information Diversity', *Journal of the American Statistical Association*, 111:516, ss. 1623–1633.
[42] Tetlock et al. (2014), 'Forecasting Tournaments', s. 291.

Som et siste tiltak identifiserte GJP en gruppe deltagere som bestod av de 2 % beste deltagerne i løpet av et år. Disse «superforecasterne» traff systematisk bedre enn alle andre eksperimentgrupper. De ble også enda bedre når de ble satt i grupper med andre superforecastere (fordelt på fem lag à tolv personer). Samtidig var det lite å hente gjennom ekstremisering av superforecasternes prediksjoner fordi de alle i utgangspunktet var svært kunnskapsrike individer og som allerede arbeidet i grupper, som betyr at det var lite informasjon «til overs».[43]

Allerede det andre året av turneringen slo snittet av prediksjonene til superforecasterne målene som IARPA hadde satt for det fjerde året av turneringen.[44] Det er derfor ikke overraskende skrevet en lang rekke artikler basert på funnene fra GJP. Siden de fleste forskerne bak prosjektet var psykologer, har imidlertid de fleste studiene vært rettet mot fagfelt som kognitiv psykologi og beslutningstaking. Fra et psykologisk perspektiv er det for eksempel interessant å vite hvilke kognitive prosesser som er involvert i prediksjon og om superforecasternes usedvanlige treffsikkerhet bare er et uttrykk for en mer generell høyere vurderingsevne («good judgment»).[45]

De psykologiske mekanismene i seg selv er imidlertid ikke like relevant i sammenheng med forsvarsplanlegging og etterretning, der det mest interessante er å forstå hvilke faktorer som bidrar til høyere treffsikkerhet og hvordan vi kan identifisere individene som treffer best. Dette er spesielt viktig i små fagmiljøer som i Norge, der enkeltpersoners prediksjoner kan få mye å si. For denne rapportens formål er det derfor GJPs studier av individuelle variasjoner i treffsikkerhet som er mest relevante å se nærmere på. De følgende underkapitlene vil derfor oppsummere GJPs funn om generelle drivere bak treffsikkerhet og hva som kjennetegnet superforecasterne.

4.1 Individuelle variasjoner

Hovedfunnet fra GJP var at det er systematiske forskjeller i individers treffsikkerhet og at denne prediksjonsevnen holder seg overraskende konsistent over tid. Et første spørsmål var derfor: Hvorfor er noen personer systematisk bedre til å predikere enn andre? Dette besvares av Mellers mfl. i en artikkel fra 2015.[46] Studien baserte seg på resultatene fra de to første årene av GJP, som inkluderte ca. 150 000 prediksjoner fra 743 deltagere på 199 spørsmål. Her målte de betydningen av tre kategorier variabler som kunne tenkes å forklare de individuelle variasjonene i treffsikkerhet: 1) disposisjonelle (individuelle forutsetninger), 2) situasjonelle (påvirkning fra omgivelsene), og 3) adferd i turneringen (hvordan deltagerne oppførte seg når de predikerte).

[43] Satopää et al. (2015), 'Modeling Probability Forecasts via Information Diversity'.
[44] Tetlock et al. (2014), 'Forecasting Tournaments', s. 292.
[45] Mellers, B., Baker, J., Chen, E., Mandel, D. og Tetlock, P. (2017), 'How generalizable is good judgment? A multi-task, multi-benchmark study', *Judgment and Decision Making*, 12:4, ss. 369–381. Her finner de for eksempel at «superforecasterne» i GJP scoret minst like godt som andre deltagere og studenter uten prediksjonserfaring på tester av konsistens, som måler i hvor stor grad en tillegger vurderinger som «en god sjanse for å ha kreft» samme sannsynlighet som «en god sjanse for regn», selv om hendelsene ofte assosieres med helt ulike prosenter.
[46] Mellers, B., Stone, E., Atanasov, P., Rohrbaugh, N., Metz, S. E., Ungar, L., Bishop, M. M., Horowitz, M., Merkle, E. og Tetlock, P. (2015a), 'The Psychology of Intelligence Analysis: Drivers of Prediction Accuracy in World Politics', *Journal of Experiment Psychology: Applied*, 21:1, ss. 1106–1115.

4.1.1 Disposisjonelle variabler

Å kunne forutsi politisk utvikling antas å forutsette en rekke evner – fra å ha grunnleggende faktakunnskap til å kunne resonnere rundt årsakssammenhenger og anvende ny informasjon. Beslutningstagere baserer seg ofte på fagfolk, nettopp fordi de antas å ha bedre forutsetninger enn andre til å kunne vurdere slike spørsmål. I GJP ble deltagernes treffsikkerhet derfor målt opp mot tre typer disposisjonelle variabler: kognitive evner, kunnskapsnivå og tenkemåter.

Kognitive evner
Den første typen disposisjonell variabel bestod av tre ulike kognitive evner, som på ulike måter ble ansett som relevante i forbindelse med politisk prediksjon:[47]

1) *Abstrakt resonneringsevne*, det vil si evnen til å trekke slutninger fra enkeltobservasjoner til mer generelle prinsipper, som er kjernen i induktiv tenkning. I prediksjonssammenheng kan det for eksempel være aktuelt å se på sammenhenger mellom et dagsaktuelt spørsmål, som sannsynligheten for et kupp i Russland, og relevante historiske tilfeller. Her må en se etter regelmessigheter, utlede hypoteser og teste dem.

2) *Kognitiv kontroll* (også kalt *kognitiv refleksjonsevne*), det vil si evnen til å unngå mentale snarveier som leder til gale svar. Tenk deg at du får følgende oppgave: «Et balltre og en ball koster 1,10 dollar. Balltreet koster 1 dollar mer enn ballen. Hvor mye koster ballen?» De fleste tenker umiddelbart at ballen koster 10 cent, men det riktige svaret er 5 cent. Denne oppgaven krever at vi tenker oss om i stedet for å gå for det svaret som faller oss inn først. Noen personer er flinkere til å unngå slike tankefeil enn andre.

3) *Tallforståelse*, det vil si evnen til å forstå tallkonsepter, som sannsynlighet. Studier har vist at selv høyt utdannede personer har vanskeligheter med relativt enkle talloppgaver, f.eks.: «Sjansen for å få en virusinfeksjon er 0,0005. Av 10 000 personer, omtrent hvor mange av dem er forventet å bli smittet?».[48] Tallforståelse antas å ha betydning for evnen til å predikere spesielt økonomiske spørsmål, som oljeprisen, siden en tallkyndig person vil lettere kunne forstå forholdet mellom dagens kurs og variasjoner over tid.

En første hypotese var derfor at individer med bedre abstrakt resonneringsevne, kognitiv kontroll og tallforståelse ville ha høyere treffsikkerhet. *Abstrakt resonneringsevne*, som ofte refereres til som et mål på flytende intelligens (eller bare intelligens), ble målt ved hjelp av en kortversjon av *Ravens Advanced Progressive Matrices* (Ravens APM).[49] Denne testen består av tolv

[47] I GJP-studiene omtales disse tre evnene ofte som aspekter av «intelligens». Bakgrunnen er at det i psykologien ofte skilles mellom «flytende» intelligens, som dreier seg om evnen til å løse nye oppgaver som i liten grad beror på tidligere læring, og «krystallisert» intelligens, som i større grad handler om på evnen til løse problemer med utgangspunkt i tidligere kunnskap. De tre kognitive evnene omtalt her er assosiert med flytende intelligens, men kognitiv refleksjonsevne omtales samtidig som noe annet enn intelligens. For enkelhets skyld brukes her «kognitive evner» om abstrakt resonneringsevne, kognitiv kontroll og tallforståelse, og «kunnskapsnivå» i stedet for krystallisert intelligens.
[48] Det riktige svaret er 5 personer. Se Lipkus, I. M., Samsa, G. og Rimer, B. K. (2001), 'General Performance on a Numeracy Scale among Highly Educated Samples', *Medical Decision Making*, 21:1, ss. 37–44.
[49] For en norsk beskrivelse av Ravens matriser, se Helland-Riise, F. og Martinussen, M. (2017), 'Måleegenskaper ved de norske versjonene av Ravens matriser [Standard Progressive Matrices (SPM)/Coloured Progressive Matrices (CPM)]', *PsykTestBarn*, 2:2.

matriseoppgaver som kan brukes uavhengig av respondentens kulturelle og lingvistiske kunnskap. *Kognitiv kontroll* ble målt ved to tester. Den første var den opprinnelige *Cognitive Reflection Test* (CRT) fra 2005, som består av tre spørsmål, hvor balltre- og balloppgaven er den første.[50] Den andre var en nyere test av det samme, som består av fire andre spørsmål, f.eks.: «Alle blomster har kronblader. Roser har kronblader. Hvis disse to påstandene er riktige, kan vi konkludere fra dem at roser er blomster?».[51] *Tallforståelse* ble målt gjennom tre oppgaver hentet fra to forskjellige tester.[52]

Mellers mfl. fant en signifikant korrelasjon mellom deltagernes treffsikkerhet og score på Ravens APM, den opprinnelige CRT-testen og den nyere, utvidede. Det var derimot ingen signifikant sammenheng mellom treffsikkerhet og tallforståelse, men reliabiliteten var usikker, fordi nesten alle deltagerne svarte riktig på alle oppgavene (2,71 av 3 riktige i snitt).

Kunnskapsnivå

Den andre typen disposisjonell variabel var deltagernes faktakunnskap om internasjonal politikk. Antagelsen var at generell politisk kunnskap er relevant for prediksjon. Hvis du for eksempel blir bedt om å anslå sannsynligheten for at FNs sikkerhetsråd vil autorisere en militær intervensjon mot Assad-regimet i Syria, vil det antagelig være en fordel å vite at rådet har fem faste medlemmer som kan legge ned veto, deriblant Russland som er alliert med regimet.

Den andre hypotesen var derfor at deltagere med høyere kunnskapsnivå ville være mer treffsikre enn dem med lavere forhåndskunnskap. Politisk kunnskapsnivå ble målt ved hjelp av to tester som ble gjennomført det første og andre året av turneringen. Her fikk deltagerne påstander som «Aserbajdsjan og Armenia har formelt avgjort sin grensekonflikt», og ble bedt om svare om de mente påstanden var riktig eller feil. Det første året bestod testen av 35 påstander og det andre året av 50 påstander. Andelen riktige svar var relativt høyt, med et gjennomsnitt på henholdsvis 82 % det første året (28,8 av 35 påstander) og 76 % det andre året (36,5 av 50 påstander).

Her fant forskerne en signifikant korrelasjon mellom begge kunnskapstestene og treffsikkerhet, til støtte for hypotesen om at politiske kunnskapsnivå predikerer treffsikkerhet.

Tenkemåter

Den tredje typen disposisjonell variabel var deltagernes tenkemåter, også kalt kognitive stiler. Kognitive stiler handler om *hvordan* folk tenker, i motsetning til *hva* de tenker på eller *hvor gode* de er. Mennesker har for eksempel forskjellige måter å behandle informasjon på. Her identifiserte Mellers mfl. tre ulike mål på kognitive stiler av relevans for treffsikkerhet:

[50] Frederick, S. (2005), 'Cognitive Reflection and Decision Making', *Journal of Economic Perspectives*, 19:4, ss. 25–42.

[51] Baron, J. Scott, S. Fincher, K. og Metz, S. E. (2015), 'Why does the Cognitive Reflection Test (sometimes) predict utilitarian moral judgment (and other things)?', *Journal of Applied Research in Memory and Cognition*, 4:3, ss. 265–284. Den fullstendige testen med alle 18 oppgaver er beskrevet i appendiks D.

[52] Den første oppgaven ble hentet fra Lipkus et al. (2001), 'General Performance on a Numeracy Scale among Highly Educated Samples', mens de to siste kom fra Peters, E., Västfjäll, D., Slovic, P., Mertz, C. K., Mazzocco, K. og Dickert, S. (2006), 'Numeracy and Decision Making', *Psychological Science*, 17:5, ss. 407–413, men det er ikke oppgitt hvilke oppgaver som er hentet fra hvilken kilde.

1) *Actively open-minded thinking* (AOMT) handler om å behandle ulike konklusjoner likt, selv om de går imot våre foretrukne svar. Personer som scorer høyt på AOMT blir mindre påvirket av eksisterende oppfatninger og er mer villige til å erkjenne at de selv kan ta feil. I en tidligere studie hadde Mellers allerede funnet at personer som scorer høyt på AOMT samler mer informasjon og at mer informasjonsinnhenting forbedrer evnen til å estimere ukjente størrelser.[53] I den grad tilgjengelig informasjon kan bidra til å forutsi fremtidige utfall, var det grunn til å anta at personer med høyere grad av AOMT også vil være bedre til å predikere enn andre.

2) *Kognitiv lukking* (*need for closure*, NFC) handler om å trekke konklusjoner raskt, ofte før alle bevis har blitt samlet, og aversjon mot tvetydighet.[54] Fordelene ved lukkethet er større handlekraft når en beslutning skal tas, men det øker også sjansen for feilslutning, fordi avgjørelser kan bli tatt for raskt eller man overser viktig informasjon. Kognitiv lukking bidrar også til at man holder fast ved oppfatninger, selv når bevisene tilsier at de er gale. I en tidligere studie hadde Tetlock funnet at eksperter med større behov for kognitiv lukking, hadde lettere for å avvise kontrafaktiske scenarioer som viste at teoriene deres var feil, mens de omfavner kontrafaktiske scenarioer som beviste at de hadde rett.[55] En antagelse var derfor at et større behov for kognitiv lukking vil være til hinder for å modellere usikkerhet ved prediksjon av hendelser i den virkelige verden.

3) *Pinnsvin- vs. revetenkning*, det vil si i hvor stor grad personer foretrekker å applisere teorier en allerede kjenner godt fra før (pinnsvin) eller om en forsøker å trekke på forskjellige vitenskapelige retninger (rev) når politiske fenomener skal forklares. Mer pinnsvinaktige personer har ofte også et større behov for kognitiv lukking.

Selv om disse tre tenkemåtene representerer forskjellige kognitive stiler, har de til felles at de handler om fordomsfrihet. En tredje hypotese var derfor at deltagere som var mer fordomsfrie ville være mer treffsikre. AOMT ble målt ved en test der deltagerne måtte oppgi hvor uenig eller enig (på en skala fra 1 til 7) de var i syv ulike påstander, f.eks. «Å endre din egen oppfatning er et tegn på svakhet». NFC ble målt ved en tilsvarende test med elleve påstander, f.eks. «Jeg liker ikke situasjoner som er usikre.». Graden av pinnsvin- vs. revetenkning ble målt utfra i hvor stor grad deltagerne vurderte seg selv som den ene eller den andre typen (på en skala fra 1 til 5).

Av disse tre forskjellige målene på fordomsfrihet var det bare AOMT som var signifikant relatert til treffsikkerhet. Dette gav bare delvis støtte til hypotesen om at fordomsfrie måter å tenke

[53] Haran, U., Ritov, I. og Mellers, B. A. (2013), 'The role of actively open-minded thinking in information acquisition, accuracy, and calibration', *Judgment and Decision Making*, 8:3, ss. 188–201.
[54] Webster, D. M. og Kruglanski, A. W. (1994), 'Individual differences in need for cognitive closure', *Journal of Personality and Social Psychology*, 67:6, ss. 1049–1062; Kruglanski, A. W. og Webster, D. M. (1996), 'Motivated closing of the mind: "Seizing" and "freezing."', *Psychological Review*, 103:2, ss. 263–283.
[55] Tetlock, P. E. (1998), 'Close-call counterfactuals and belief-system defenses: I was not almost wrong but I was almost right', *Journal of Personality and Social Psychology*, 75:3, ss. 639–652.

på predikerer treffsikkerhet. Det betød også at skillet mellom pinnsvin og rever hadde mindre betydning for treffsikkerheten enn i EPJ.

4.1.2 Situasjonelle variabler

Den andre kategorien variabler som kunne tenkes å forklare variasjonene i deltagernes treffsikkerhet, var påvirkningen de fikk fra omgivelsene. Som beskrevet i oppsummeringen av vinneroppskriften til GJP, viste det seg at opplæring i probabilistisk tenkning og det å bli satt i gruppe med andre deltager bidro til å øke treffsikkerheten sammenlignet med å predikere alene.[56]

En fjerde hypotese som GJP ønsket å teste, var derfor om de disposisjonelle variablene ville øke treffsikkerheten utover disse situasjonelle variablene. Denne hypotesene ble analysert ved en multippel regresjonsanalyse der funnet var at kognitive evner og politisk kunnskapsnivå, men ikke fordomsfrihet, bidro til økt treffsikkerhet utover opplæring og gruppearbeid.

4.1.3 Adferdsvariabler

En tredje og siste variabelkategori som kunne tenkes å påvirke treffsikkerheten, var hvordan deltagerne *oppførte* seg i selve turneringen. Mengdetrening regnes som avgjørende for prestasjonsevnen innenfor mange områder, som sport og musikk.[57] Tidligere studier har også vist at personer med et såkalt *growth mindset* – det vil si at en anser læring og oppnåelse som ferdigheter som kan dyrkes – har større sannsynlighet for å prestere godt enn personer med et *fixed mindset* – der evner bare anses som medfødte («Jeg er dårlig i matematikk»).[58] Personer med et *growth mindset* liker utfordringer og klarer oftere å forbedre evnen sine, mens personer med et *fixed mindet* har lettere for å gi opp når det blir vanskelig. Betydningen av trening og den positive effekten av *growth mindset* kunne derfor også tenkes å gjelde innenfor prediksjon.

En femte hypotese var derfor at en adferd som reflekterte et *growth mindset* ville predikere økt treffsikkerhet utover de disposisjonelle og situasjonelle variablene. Adferd ble målt på tre måter: antall spørsmål deltagerne svarte på, antall prediksjoner per spørsmål og tiden de brukte per spørsmål. Høyere verdier ble vurdert som uttrykk for et sterkere *growth mindset*.

Denne antagelsen ble også analysert ved multippel regresjonsanalyse, der disse tre typene adferd ble målt opp mot kognitive evner, kunnskapsnivå, tenkemåter, opplæring og gruppearbeid. Til støtte for hypotesen korrelerte antallet prediksjoner og tiden brukt per spørsmål med treffsikkerhet, mens antallet spørsmål de svarte på ikke gjorde det. Faktisk var antallet prediksjoner per spørsmål den variabelen som korrelerte sterkest med treffsikkerheten, mens tid brukt per spørsmål var den nest sterkeste.[59]

---

[56] Mellers et al. (2014), 'Psychological strategies for winning a geopolitical forecasting tournament'.
[57] Ericsson, K. A., Krampe, R. T. og Tesch-Romer, C. (1993), 'The role of deliberate practice in the acquisition of expert performance', *Psychological Review*, 100:3, ss. 363–406.
[58] 'Growth mindset', *Store norske leksikon*.
[59] For mer om sammenhenger mellom treffsikkerhet og oppdatering av prediksjoner, se Atanasov, P., Witkowski, J., Ungar, L., Mellers, B. og Tetlock, P. (2020), 'Small steps to accuracy: Incremental belief updaters are better forecasters', *Organizational Behavior and Human Decision Processes*, 160, ss. 19–35.

4.1.4 Oppsummering

Tabell 4.1 viser korrelasjonene mellom treffsikkerheten og alle individuelle variabler i GJP. Høyere individuell treffsikkerhet var assosiert med bedre score på kognitive evner, politisk kunnskapsnivå og til dels fordomsfri tenkning. Disse egenskapene økte treffsikkerheten utover situasjonelle variabler. Det var likevel hvor ofte deltagerne oppdaterte sine prediksjoner, og hvor lang tid de brukte på spørsmålene, som korrelerte sterkest med treffsikkerheten.

	Mål	**Korrelasjon**	**t(741)**
Kognitive evner	Abstrakt resonneringsevne (Ravens)	**−0.23**	−6.38
	Kognitiv kontroll (CRT med 3 oppg.)	**−0.15**	−4.17
	Kognitiv kontroll (utvidet CRT med 4 oppg.)	**−0.14**	−3.56*
	Tallforståelse	−0.09	
Tenkemåter	Actively open-minded thinking (AOMT)	**−0.10**	−2.51**
	Need for closure (NFC)	0.03	
	Pinnsvin- vs. revetenkning	0.09	
Kunnskaps-nivå	Politisk kunnskapsnivå (1. år)	**−0.18**	−4.85
	Politisk kunnskapsnivå (2. år)	**−0.20**	−5.06***
Situasjon	Opplæring	**−0.17**	−4.56
	Gruppearbeid	**−0.30**	−8.55
Adferd	Antall prediksjoner per spørsmål	**−0.49**	−15.29
	Antall spørsmål besvart	0.07	
	Tid brukt per spørsmål	**−0.30**	−8.28****

*Tabell 4.1 Korrelasjoner mellom treffsikkerhet (standardisert Brier-score) og alle individuelle variabler. Fet skrift indikerer en signifikant forskjell på .001-nivå. * t(599), * t(742), *** t(648), **** t(694). Gjengitt med tillatelse.*[60]

For å undersøke sammenhengene mellom variablene nærmere, benyttet Mellers mfl. *Structural Equation Modeling* (SEM). Her ble antall svar per spørsmål identifisert som en mellomliggende variabel mellom kognitiv evne og treffsikkerhet, mellom kunnskapsnivå og treffsikkerhet og mellom gruppearbeid og treffsikkerhet, mens tid brukt per spørsmål var en mellomliggende variabel mellom gruppearbeid og treffsikkerhet. Disse årsakssammenhengene kan tolkes ulikt: Deltagere med høyere kunnskapsnivå og kognitiv evne kan ha likt oppgavene mer, og dermed deltatt mer aktivt. Alternativt kan deltagerne ha blitt mer kunnskapsrike etter hvert som de ble mer engasjert. De som arbeidet i grupper kan ha blitt motivert av ønsket om å gjøre det godt for gruppens skyld, som også kan ha bidratt til hyppigere oppdateringer og høyere treffsikkerhet.

[60] Dette er en gjengivelse av tabell 2 i Mellers et al. (2015a), 'The Psychology of Intelligence Analysis', s. 8.

4.2 Superforecastere

I en annen artikkel fra 2015, gikk Mellers mfl. nærmere inn på hva som kjennetegnet de aller beste deltagerne – *superforecasterne*.[61] Disse bestod av de 2 % beste av over 1700 deltagere. Dette datagrunnlaget baserte seg på rundt 350 spørsmål fra de tre første årene av GJP.

De første superforecasterne ble identifisert etter at det første året av turneringen var over. Deltagerne ble rangert ut fra treffsikkerheten, og de 5 beste deltagerne innenfor prosjektets 12 ulike eksperimentelle betingelser, til sammen 60 deltagere, ble plukket ut som superforecastere. Disse ble så fordelt på 5 nye superforecaster-lag med 12 medlemmer hver. I tillegg til å bli satt i grupper fikk alle som ikke hadde fått det også opplæring i kognitive fallgruver, siden disse to situasjonelle tiltakene hadde vist seg å øke treffsikkerheten det første året.

For å kunne måle den relative treffsikkerheten til superforecasterne, ble resten av deltagerne i GJP delt inn i to grupper:

1) *Top-team individuals*, som bestod av de nest beste deltagerne som var satt i grupper. De var altså svært gode, men nådde ikke opp til superforecaster-kriteriet. Hensikten med å sammenligne superforecasterne med disse var å måle forskjellen mellom treffsikkerheten til elitegrupper og vanlige grupper.

2) *Alle andre deltagere*, som bestod av rundt 1500 personer.

Både superforecasterne og top-team individuals traff langt bedre enn resten av deltagerne, men superforecasterne var likevel systematisk bedre enn de nest beste over tid. Hvis superforecasternes høye treffsikkerhet det første året bare hadde vært flaks, burde treffsikkerheten deres ha falt tilbake mot snittet i det andre og tredje året (regresjon mot middelverdien). Tvert imot scoret superforecasterne bedre både det andre og tredje året av turneringen enn det første, mens top-team individuals og andre deltagere traff gradvis dårligere i påfølgende år og forskjellen dem imellom ble mindre.[62] 70 % av superforecasterne forble derimot blant de 2 % året etter, som betyr at sannsynligheten for at prediksjonsevnen deres bare var tilfeldig er svært liten.

Superforecasterne traff også signifikant bedre enn de to andre gruppene når en tok høyde for når deltagerne predikerte og tiden deltagerne brukte på spørsmålene. Dette ble undersøkt ved å begrense utvalget til: 1) bare prediksjoner som ble gjort den første dagen et spørsmål ble lansert, og 2) bare prediksjoner som ble levert ila. fire minutter fra tidspunktet deltagerne fikk se spørsmålet til prediksjonene var levert, som gav liten tid til å lese seg opp. Selv med relativt liten tid og innsats var superforecastere bedre til å predikere.

[61] Oppsummeringen er hovedsakelig basert på Mellers, B., Stone, E., Murray, T., Minster, A., Rohrbaugh, N., Bishop, M., Chen, E., Baker, J., Hou, Y., Horowitz, M., Ungar, L. og Tetlock, P. (2015b), 'Identifying and Cultivating Superforecasters as a Method of Improving Probabilistic Predictions', *Perspectives on Psychological Science*, 10:3, ss. 267–281.
[62] Mellers et al. (2015b), 'Identifying and Cultivating Superforecasters', s. 270.

Samlet sett viste det seg at det å samle superforecasterne sammen på egne elitelag økte treffsikkerheten mer enn det opplæring i probabilistisk tenkning og gruppearbeid gjorde til sammen. Superforecasterne skal også ha truffet 30 % bedre enn et prediksjonsmarked med amerikanske etterretningsanalytikere som hadde tilgang på gradert informasjon.[63] Dette var omtrent den samme marginen som superforecasterne slo GJPs eget prediksjonsmarked med.[64]

For å undersøke hva som gjorde at superforecasterne var så supre, sammenlignet Mellers mfl. hvordan superforecasterne scoret sammenlignet med top-team individuals og resten av deltagerne på de samme variablene som i studien av individuelle variasjoner (se tabell 4.2).

	Mål	**Super-forecastere**	**Top-team individuals**	**Resten**	**Kon-troll**
Kognitive evner	Ravens matriser (0–12)	9.13	**8.26**	**7.75**	
	Shipley–2 Abstraction (0–25)	20.09	**18.58**	**18.49**	
	Kognitiv kontroll, opprinnelig CRT (0–3)	2.78	**2.46**	**2.26**	
	Kognitiv kontroll, utvidet CRT (0–18)	16.64	**15.39**	**14.56**	
	Tallforståelse (0–4)	3.67	**3.33**	**3.19**	
Tenke-måter	Motivasjon? Være blant de beste (1–7)	5.60	**4.86**	**4.81**	
	Kognitiv motivasjon (1–7)	5.97			
	Actively open-minded thinking (1–7)	6.01	5.97	**5.89**	
	Tro på skjebnen (0–7)	2.65			**3.17**
	Close calls	.16			**.40**
Kunn-skaps-nivå	Politisk kunnskapsnivå 1. år (0–35)	29.59	29.37	**28.66**	
	Politisk kunnskapsnivå 2. år (0–50)	38.45	**37.30**	**36.29**	
	Politisk kunnskapsnivå 3. år (0–55)	32.20	31.23	**31.12**	
	Shipley–2 Vocabulary (0–40)	37.50	**36.89**	**36.79**	
Oppga-vespesi-fikke fer-dig-heter	*Scope sensitivity*	.22			**0.12**
	Granularity (antall unike sannsynlighetsestimater)	57	**29**	**30**	
Adferd	Antall spørsmål besvart (1. år)	76	**65**	**57**	

[63] Tetlock, P. E., Mellers, B. A. og Scobilic, J. P. (2017), 'Bringing probability judgments into policy debates via forecasting tournaments', *Science*, 355:6324, ss. 481–483.
[64] 'Edge Master Class 2015'.

	Antall spørsmål besvart (2. år)	116	**84**	**82**	
	Antall spørsmål besvart (3. år)	81	**52**	**60**	
	Gj.snitt antall prediksjoner per spørsmål (1. år)	2.77	**1.51**	**1.43**	
	Gj.snitt antall prediksjoner per spørsmål (2. år)	5.64	**2.15**	**1.79**	
	Gj.snitt antall prediksjoner per spørsmål (3. år)	6.70	**5.14**	**2.92**	
	Gj.snitt antall nyhetsartikler lest (2. år)	187.31	**45.72**	**24.89**	
	Gj.snitt antall nyhetsartikler lest (3. år)	344.73	**63.12**	**89.80**	
Situasjon (elite-grupper vs. vanlige grupper)	Gj.snitt antall kommentarer (2. år)	262.23	**51.88**		
	Gj.snitt antall kommentarer (3. år)	622.89	**112.26**		
	Gj.snitt antall ord per kommentar (2. år)	36.62	**28.49**		
	Gj.snitt antall ord per kommentar (3. år)	31.66	**24.80**		
	Gj.snitt antall poster (2. år)	36.13	**2.25**		
	Gj.snitt antall poster (3. år)	43.64	**4.94**		
	Gj.snitt antall nyhetsartikler delt (2. år)	91.57	**9.16**		
	Gj.snitt antall nyhetsartikler delt (3. år)	181.93	**28.61**		
	Gj.snitt % kommentarer med spørsmål (2. år)	0.47	**0.19**		
	Gj.snitt % kommentarer med spørsmål (3. år)	0.32	0.27		
	Gj.snitt % svar (2. år)	7.29	**2.59**		
	Gj.snitt % svar (3. år)	6.54	**2.99**		
	Consensus rate	−0.06	**0.05**	**0.06**	

Tabell 4.2 Sammenligning av superforecastere og kontrollgrupper. Fet skrift hos top-team individuals og resterende deltagere indikerer en signifikant forskjell sammenlignet med superforecastere på .01-nivå. Gjengitt med tillatelse.[65]

[65] Dette er en gjengivelse av tabell 2 i Mellers et al. (2015b), 'Identifying and Cultivating Superforecasters', s. 274.

4.2.1 Disposisjonelle variabler

Kognitive evner
Foruten den samme testen av abstrakt resonneringsevne (Ravens) og den opprinnelige CRT-testen, ble den utvidede testen av kognitiv kontroll med fire oppgaver erstattet av en større versjon med 18 oppgaver (Baron et al 2015) ved begynnelsen av det tredje året av turneringen.

I tillegg ble det introdusert en helt ny *Abstraction Test* fra *Shipley Institute of Living Scale 2 (Shipley-2*).[66] Her måles abstrakt resonneringsevne ved 25 oppgaver der deltagerne må fullføre sekvenser som «*white/black*, *short/long*, *down*/…» og «*oh/ho*, *rat/tar*, *mood*/…)». For deltagere som deltok alle tre årene ble testresultatene standardisert for å sikre sammenlignbare verdier.

Testen av tallforståelse ble også utvidet fra tre til fire oppgaver, siden reliabiliteten hadde vært under tvil, men oppgavene ble hentet fra samme kilder.[67] På alle kognitive evnetestene scoret superforecasterne signifikant høyere enn både top-team individuals og resten av deltagerne.

Kunnskapsnivå
Det tredje året ble det gjennomført en ytterligere test av politisk kunnskap, denne gang med 55 påstander. I tillegg ble det introdusert en ny *Vocabulary Test*, som også var hentet fra *Shipley–2*. Denne representerte et annet mål på tilegnet kunnskap, der deltagerne fikk 40 oppgaver hvor de måtte finne ut hvilke ord som lå nærmest hverandre i mening. For eksempel fikk de ordet «*large*» og måtte velge hvilket av ordene «*red – big – silent – wet*» som lignet mest.

Superforecasterne scoret høyere enn top-team individuals på de politiske kunnskapstestene alle årene av turneringen, men denne forskjellen var bare signifikant på -.01-nivå det andre året. Superforecasternes score var imidlertid signifikant høyere enn resten av deltagerne alle tre årene. Superforecasterne var også signifikant bedre enn begge kontrollgruppene på den nye vokabulartesten.

Tenkemåter
Av de tre kognitive stilene som ble målt i studien av individuelle variasjoner, var det bare *actively open-minded thinking* (AOMT) som korrelerte med treffsikkerhet. I tråd med dette funnet scoret superforecasterne litt høyere enn resten også her, men forskjellen mellom dem og top-team individuals var ikke signifikant på .01-nivå. De to øvrige testene av tenkemåter som ikke korrelerte med treffsikkerhet i studien av individuelle variasjoner (kognitiv lukking og pinnsvin- vs. revetenkning), ble ikke inkludert i denne superforecaster-studien.

Det ble imidlertid introdusert fire nye mål på kognitive stiler. Den første målte deltagernes motivasjon for å delta. Alle deltagerne ble spurt: «Hvorfor valgte du å delta i turneringen?». Her uttrykte superforecasterne et betydelig høyere ønske om å «havne blant de beste» enn både top-

[66] Shipley, W. C., Gruber, C. P., Martin, T. A. og Klein, A. M. (2009), *Shipley-2 Manual* (Western Psychological Services).
[67] Se fotnote 6 i appendiks 1d i Friedman, J. A. (2019), *War and Chance: Assessing Uncertainty in International Politics* (Oxford University Press).

team individuals og resten. Den andre nye testen var av kognitiv motivasjon (*need for cognition*), som handler om folks behov for og glede av å engasjere seg i mentalt krevende aktiviteter.[68] Her måles ikke individers evner, men viljen til å engasjere seg i oppgaver som krever dypere tenkning og til å jobbe med den kapasiteten en har. Folk med høyere kognitiv motivasjon setter større pris på diskusjoner og problemløsningsoppgaver, mens folk med lavere score har lettere for å ta mentale snarveier. Her scoret superforecasterne relativt høyt, men det er ikke oppgitt tall for de to andre gruppene.

Den tredje nye testen målte deltagernes tro på skjebnen (*belief in fate*).[69] I prediksjon er det en grunnleggende forskjell mellom å anta en gudommelig orden, der alt skjer av en grunn, og å ha en vitenskapelig, probabilistisk tilnærming der fremtidige hendelser vurderes ut fra sannsynligheter. Graden av tro på skjebnen ble målt ut fra hvor enige deltagerne var i pro-probabilistiske påstander, som «Tilfeldigheter er ofte en faktor i våre personlige liv», og hvor uenige de var i pro-skjebne påstander, som «Hendelser utspiller seg etter Guds plan».

Resultatene viste at den gjennomsnittlige amerikanske voksne befolkningen lå omtrent midt på skalaen, mens bachelorstudenter ved et ledende universitet i USA scoret noe lavere på tro på skjebnen. Vanlige deltagere i GJP scoret enda litt lavere, mens superforecasterne scoret lavest. Det ble også funnet en korrelasjon mellom høyere treffsikkerhet og lavere tro på skjebnen; altså, at jo mer probabilistisk deltagerne tenkte, jo mer treffsikre var de.

Den fjerde nye testen av tenkemåter var av deltagernes reaksjoner på hendelser som *bare nesten* skjedde/ikke skjedde (*close calls*). En vanlig reaksjon når du finner ut at noe bare så vidt skjedde, f.eks. et par som møtte hverandre på en høyst uvanlig plass å være, i motsetning til et par som traff hverandre en helt vanlig plass å møtes, er å attribuere hendelsene som bare så vidt skjedde til skjebnen («det var ment å skje»). I et eget eksperiment, der tre grupper deltagere fikk lese enten *close-call*-historier eller ikke-*close-call*-historier, var superforecasterne mest tilbøyelig til å attribuere hendelsene som så vidt skjedde til tilfeldigheter og minst tilbøyelig til å attribuere dem til skjebnen.

Oppgavespesifikke ferdigheter

I tillegg til disse tre typene disposisjonelle variabler, målte Mellers mfl. også to nye, mer spesifikke, evner knyttet til prediksjon som oppgave.

Den første er *scope insensitivity* (også kjent som *scope neglect*), som omhandler hvordan vi sliter med å forstå omfanget av størrelser, spesielt når det er snakk om store tall. Dette gjør at vi også har vanskeligheter med å justere våre vurderinger proporsjonalt i forhold til størrelsen eller omfanget på det vi blir spurt om, som f.eks. forskjellen på hjelpetiltak som kan redde få eller mange.

[68] Reierth, M. G. og Tronstad, J. (2015), 'Effektive team: Kognitiv motivasjon og maksimerings betydning for hvordan team arbeider og presterer' (Norges handelshøyskole).
[69] For mer om denne testen, se Tetlock og Gardner (2015), *Superforecasting*, ss. 147–152, og Mellers et al. (2015b), 'Identifying and Cultivating Superforecasters', s. 273

I et eksperiment ble respondenter bedt om å oppgi hvor mye de var villige til å betale for beskytte trekkfugler fra oljesøl. Respondentene ble delt inn i tre grupper, der de ble fortalt at det var hhv. 2000, 20 000 eller 200 000 fugler som ble påvirket av oljesøl årlig. Når de ble spurt om hvor mye de var villige til å betale for å beskytte fuglene, oppgav imidlertid alle gruppene omtrent det samme ($80–$90), selv om antallet fugler som ble rammet varierte mye mer.

I prediksjonssammenheng kan et spørsmåls omfang variere med hensyn til for eksempel tidsperspektivet. Sannsynligheten for at en bestemt hendelse kan skje innenfor et kort tidsperspektiv kan umulig være større enn sannsynligheten for at det skjer innenfor et lengre perspektiv. Ta for eksempel følgende spørsmålspar: «Vil det syriske regimet falle innen 2021?» og «Vil det syriske regimet falle innen 2023?». Rent logisk må sannsynligheten for at det syriske regimet faller i løpet av ett år være lik (eller trolig mindre) enn for at det skjer i løpet av tre, siden det første året er en underkategori av de neste tre.

For å måle i hvor stor grad deltagerne hadde *scope sensitivity*, fikk de fire spørsmålspar, der omfanget varierte på ulike måter. Deretter målte de forskjellene i sannsynlighetsestimatene på de mest og minst sannsynlige hendelsene til superforecasterne og en kontrollgruppe. Større forskjell innebærer større sensitivitet (selv om dette ikke nødvendigvis betyr større treffsikkerhet). På tre av fire spørsmålspar var superforecasterne mer sensitive for omfanget enn resten.

Deltagerne ble også målt på sensitivitet for ankringseffekten, som er en kognitiv bias der vi justerer våre estimater ut fra et bestemt tall (ankeret) vi får presentert først (f.eks. prisantydningen på en bolig). Et dårlig anker kan imidlertid gjøre at vi ikke justerer nok til å treffe godt. Studier har også vist at vi påvirkes av tall som ikke har noe med saken å gjøre.

For å måle ankringseffekten fikk deltagerne et spørsmål om hvor mange prosent de trodde verdensøkonomien ville vokse det neste året. I forkant av dette spørsmålet fikk de ulike deltagere et annet spørsmål med to forskjellige, men ubetydelige ankertall. Den første gruppen ble spurt om de trodde verdensøkonomien ville vokse «mer enn 2,8 %», mens en annen ble spurt om de trodde den ville vokse «mer enn 3,3 %». Variasjonen i spørsmålsformuleringene skal imidlertid ikke ha noe å si for din vurdering av hvor mange prosent økonomien vil vokse året etter, men slike ankertall har likevel en tendens til å påvirke våre estimater. Som ved andre mål på *scope sensitivity* var superforecasterne mindre utsatt for ankringseffekten. Superforecasternes svar på verdensøkonomiens vekst (hhv. 3,2 % og 3,4 % på de to spørsmålsvariantene) var mindre påvirket av ankrene (2,8 % og 3,3 %) enn kontrollgruppens svar (hhv. 2,7 % og 3,1 %).

Den andre oppgavespesifikke ferdigheten var *forecasting granularity*, det vil si hvor «finkornet» sannsynlighetsvurderingene var. Det å tallfeste sannsynligheten av fremtidige hendelser er ikke enkelt, spesielt ikke når spørsmålene er kvalitative av natur. Noen personer bryter imidlertid sannsynlighetsskalaen (fra 0 % til 100 %) ned i flere distinksjoner enn andre. Hvis du bruker svar som 23 %, 27 %, 47 %, 53 %, 74 % og 78 % er dine prediksjoner mer finkornede enn om du bare bruker 25 %, 50 % og 75 %.

I GJP ble deltagernes *forecasting granularity* målt ved å telle antallet unike sannsynlighetsvurderinger deltagerne brukte i løpet av turneringen. I snitt brukte superforecasterne dobbelt så

mange sannsynlighetsestimater som både top-team individuals og resten av deltagerne. Deretter fant en andelen prediksjoner som var multipler av 10 % (dvs. 10 %, 20 %, 30 %, osv.), 5 % (men ikke multipler av 10 %) og 1 % (men ikke multipler av 10 % eller 5 %). Top-team individuals og alle andre deltagere var de mest sannsynlige gruppene til å gjøre prediksjoner som var delelige på 10 % (10 %, 20 %, 30 %, osv.), mens superforecasterne var de mest sannsynlige til å gjøre prediksjoner som bare var delelige på 1 % (f.eks. 17 %, 28 % og 83 %).

Superforecasterne var med andre ord mer finkornede i sine prediksjoner. Mer finkornede prediksjoner er imidlertid ikke nødvendigvis relatert til høyere prediksjonsevne, hvis ikke mer nyanserte prediksjoner også bidrar til høyere treffsikkerhet. Deltagernes prediksjoner ble derfor avrundet til nærmeste 5 %, 10 % og 33 %, for å se om dette påvirket treffsikkerheten. Hvis treffsikkerheten ble dårligere etter å rundet av til nærmeste 5 %, betyr det at mer finkornede prediksjoner enn de 21 ulike distinksjonene som denne avrundingen tillater, inneholdt informasjon som økte treffsikkerheten. Tilsvarende ville en lavere treffsikkerhet etter avrunding fra 10 % til 33 % bety at prediksjonsverdig informasjon gikk tapt når sannsynlighetsestimatene ble redusert fra 11 til 4 distinksjoner. For superforecasterne innebar en avrunding til nærmeste 10 % en betydelig dårligere treffsikkerhet, mens for de to andre gruppene falt ikke treffsikkerheten før avrunding til nærmeste 33 %. Dette betyr at superforecasternes mer finkornede prediksjoner bidro til å øke treffsikkerheten.

Dette funnet har forskerne bak GJP brukt til å argumentere for at etterretningsmiljøer bør erstatte brede, kvalitative beskrivelser av sannsynlighet (f.eks. «mest sannsynlig») med tallfestede sannsynlighetsvurderinger, fordi denne spesifiseringen i seg selv kan bidra til økt treffsikkerhet, og fordi distinksjoner i sannsynlighet kan ha mye å si for store politiske beslutninger.[70]

4.2.2 Situasjonelle variabler

I undervisningssammenheng har studier vist at elever som arbeider i grupper med andre på samme evnenivå, motiverer hverandre mer, liker oppgavene bedre og lærer raskere – og at denne effekten er størst for de flinkeste.[71] I lys av denne forskningen ønsket Mellers mfl. å måle hvorvidt *elitegrupper* med bare superforecastere skilte seg fra *vanlige* grupper bestående av top-team individuals.

For å undersøke hvorvidt elitegrupper oppførte seg annerledes enn vanlige grupper, ble deltagernes interaksjoner målt på flere måter. Først telte de hvor mange kommentarer deltagerne postet til spørsmålene som ble stilt og på turneringens generelle forum. Superforecasterne postet rundt fem ganger flere kommentarer til turneringsspørsmålene, og superforecasternes kommentarer var rundt en tredel lengre. Superforecasterne postet også rundt ti ganger så mange kommentarer på det vanlige forumet enn top-team individuals.

[70] Friedman, J. A., Baker, J. D., Mellers, B. A., Tetlock, P. E. og Zeckhauser, R. (2018), 'The Value of Precision in Probability Assessment: Evidence from a Large-Scale Geopolitical Forecasting Tournament', *International Studies Quarterly*, 62:2, ss. 410–422; Friedman (2019), *War and Chance*.
[71] Mellers et al. (2015b), 'Identifying and Cultivating Superforecasters', viser til Epple, D. og Romano, R. (2011), 'Peer effects in education: A survey of the theory and evidence', *Handbook of Social Economics*, 1, ss. 1053–1163.

For å hjelpe deltagerne med å finne relevante artikler fra troverdige og relevante kilder, publiserte GJP nyhetsartikler og kronikker på turneringens nettportal. Hvor ofte deltagerne delte disse artiklene med andre, ble brukt som et tredje mål på gruppesamarbeid. Superforecasterne delte ti ganger flere nyhetsartikler enn top-team individuals det andre året, og seks ganger flere artikler det tredje året.

Andelen setninger som inneholdt spørsmålstegn var også dobbelt så høy blant superforecasterne enn blant top-team individuals, og superforecasterne svarte på en mye større andel av hverandre sine spørsmål. Superforecasterne utviste således større interesse for lagkameratenes kunnskap og hjalp hverandre mer enn det top-team individuals gjorde. Et høyere antall nyhetsartikler delt, antall kommentarer med spørsmålstegn og antall svar på andres spørsmål var alle variabler som hver for seg korrelerte med høyere treffsikkerhet.

En siste variabel var hvor raskt superforecasterne oppnådde enighet om sine prediksjoner (*consensus rate*). Velfungerende grupper vil kanskje være uenige i starten, men likevel kunne oppnå konsensus etter hvert. For å undersøke dette ble den daglige variasjonen i hver deltagers prediksjoner målt opp mot spørsmålets varighet. Mens prediksjonene til top-team individuals og alle andre deltagere spriket stadig mer mot slutten av spørsmålsperioden, ble superforecasterne gradvis mer enige. Det antas derfor at ved å dele flere artikler, stille flere spørsmål og hjelpe hverandre mer, nådde superforecasterne raskere konsensus om prediksjonene, som i tillegg var mer treffsikre.

4.2.3 Adferdsvariabler

I tråd med funnet fra den forrige artikkelen om individuelle variasjoner i treffsikkerheten, utviste superforecasterne også et enda tydeligere *growth mindset* enn de andre deltagerne i GJP.

I alle årene av turneringen svarte superforecasterne på flere spørsmål enn både top-team individuals og alle andre. Det første året svarte de på 25 % flere spørsmål, men i de påfølgende årene økte dette til rundt 40 % flere spørsmål enn de andre gruppene. Superforecasterne oppdaterte også sine prediksjoner oftere enn både top-team individuals og alle andre. I tillegg leste superforecasterne langt flere nyhetsartikler som ble postet på turneringsnettsiden (utover å dele dem).

Konklusjonen var derfor at superforecasterne var mer dedikert til turneringen og til å utvikle sin egen prediksjonsevnene enn resten. I artikkelen om superforecasterne rapporteres imidlertid ikke tiden superforecasterne brukte på per spørsmål.

4.2.4 Oppsummering

Det viktigste funnet fra GJP var at superforecasterne skilte seg systematisk ut fra alle andre. De scoret signifikant høyere på alle mål på kognitive evner, kunnskapsnivå og tenkemåter enn resten av deltagerne i GJP. Superforecasterne scoret samtidig ikke veldig mye høyere på intelligens- og kunnskapstestene enn resten av deltagerne i GJP i et større perspektiv. Vanlige deltagere scoret høyere enn 70 % av den gjennomsnittlige amerikanske befolkningen, mens superforecasterne scoret høyere enn rundt 80 %.[72] Den største forskjellen var altså mellom gjennomsnittsbefolkningen og deltagerne i GJP generelt. Superforecasterne var litt bedre, men en trenger altså ikke å være Mensa-medlem og ha en PhD fra Harvard for å være superforecaster.

Tabell 4.3 oppsummerer hvordan hver av disse disposisjonelle variablene korrelerte med treffsikkerhet, basert på resultatene fra de tre første årene av turneringen.

	Variabel	**Korrelasjon**	**t(1774)**	**p**
Kognitive evner	Abstrakt resonneringsevne (Ravens)	−0.18	−7.70	<.001
	Kognitiv kontroll (3 oppg.)	−0.16	−6.82	<.001
	Kognitiv kontroll (18 oppg.)	−0.23	−9.95	<.001
	Tallforståelse	−0.16	−6.82	<.001
	Abstrakt resonneringsevne (Shipley-2)	−0.22	−9.49	<.001
Tenke-måter	Actively open-minded thinking (AOMT)	−0.12	−5.09	<.001
	Motivasjon? Være blant de beste	−0.11	−4.66	<.002
	Kognitiv motivasjon	−0.07	−2.95	<.001
Kunn-skapsnivå	Politisk kunnskapsnivå (1. år)	−0.12	−5.09	<.001
	Politisk kunnskapsnivå (2. år)	−0.18	−7.70	<.001
	Politisk kunnskapsnivå (3. år)	−0.14	−5.95	<.001
	Vokabular (Shipley-2)	−0.09	−3.80	<.001
Adferd	Gj.snitt antall nyhetsartikler lest	−0.18	−7.70	<.001
Situasjon	Gj.snitt antall artikler delt	−0.20	−8.53	<.001
	Gj.snitt antall kommentarer med spørsmål	−0.18	−7.68	<.001
	Gj.snitt antall svar på spørsmål	−0.18	−7.70	<.001

Tabell 4.3 Korrelasjoner med treffsikkerhet. Gjengitt med tillatelse.[73]

[72] Tetlock og Gardner (2015), *Superforecasting*, s. 109.
[73] Dette er en gjengivelse av tabell 3 i Mellers et al. (2015b), 'Identifying and Cultivating Superforecasters', s. 275.

Den utvidete testen av kognitiv kontroll med 18 oppgaver og den helt nye abstraksjonstesten var de to målene som korrelerte sterkest med treffsikkerhet. Etter å ha utvidet testen av tallforståelse fra tre til fire oppgaver, korrelerte nå også denne variabelen med treffsikkerhet, som forsterker funnet om at kognitive evner har stor betydning. Ellers var korrelasjonene svært lik studien av individuelle variasjoner fra de to første årene (se tabell 4.1).

Samtidig var det én adferdsvariabel og tre situasjonelle variabler, som målte interaksjonen mellom deltagere som arbeidet i gruppe, som også korrelerte med høyere treffsikkerhet. Den første var antallet nyhetsartikler lest, og de tre andre var antall artikler deltagere delte med andre lagmedlemmer, antall spørsmål stilt og antall svar gitt på andres kommentarer.

Mellers mfl. konkluderte med at superforecasterne i GJP ble delvis «identifisert», basert på et sett med kognitive evner og stiler som er relevant for prediksjonsevne, og delvis «skapt», gjennom å tilrettelegge for forbedring av prediksjonsevnen. Turneringen gav en mulighet til å kunne trene prediksjonsevnen. De beste deltagerne brukte denne muligheten ved å svare på flere spørsmål, oppdatere sine prediksjoner og lese flere nyhetsartikler. Å sette deltagere sammen i grupper økte treffsikkerheten ytterligere, spesielt i elitegrupper.

Samtidig skilte superforecasterne seg fra de andre deltagerne allerede før turneringen begynte. Alle superforecasterne lignet mest på revene fra EPJ. De scoret høyere på tester av kognitive evner og stiler, men også på ønsket om å vinne, interesse for dypere tenkning og hadde generelt en mer vitenskapelig tilnærming til det å vurdere fremtidige hendelser. De var også mer sensitive for omfanget av spørsmålene de fikk og var mer nyanserte i sine prediksjoner.

Mens politisk kunnskap kan tilegnes og adferd motiveres, anses kognitive evner som mer faste. Tenkemåter kan endres, men fordomsfri tenkning, motivasjonen for å vinne og appetitten for dypere tenkning hadde relativt mindre å si for treffsikkerheten enn de fleste kognitive evnene og deltagernes dedikasjon. Det er derfor ikke grunn til å tro at alle kan bli superforecastere om de bare øver nok, men de fleste kan trolig bli bedre til å predikere gjennom tilrettelegging for dette.

5 Pågående studier

I 2014 ble det etablert en kommersiell spin-off av GJP, *Good Judgment Inc.*, som tilbyr kurs og prediksjonstjenester til bedrifter. Dette selskapet står også bak en åpen prediksjonsturnering, *GJ Open*, der alle som vil kan registrere seg og svare på spørsmål om alt fra geopolitikk til basketball.[74] Ifølge personvernerklæringen kan resultatene herfra brukes til forskning, men det er per i dag ikke publisert noen studier basert på turneringene på *GJ Open*.

IARPA har arrangert flere nye aktiviteter som ikke er avsluttet ennå.[75] I 2016 lanserte IARPA en «hybrid» prediksjonsturnering, *Hybrid Forecasting Competition* (HFC), der målet var å forbedre treffsikkerheten på geopolitiske prediksjoner ved å kombinere styrkene til mennesker og maskiner.[76] Tetlock omtaler dette som «neste generasjons prediksjonsturneringer».[77] Hypotesen er at maskiner vil være bedre til å predikere på spørsmål hvor det finnes lange, kvantifiserbare tidsserier, slik som bruttonasjonalprodukt, mens mennesker vil være bedre på spørsmål med få kvantifiserbare historiske data, som om det blir krig mellom USA og Iran den neste måneden. Parallelt med hybridturneringen har IARPA også arrangert to runder av en ny, vanlig prediksjonsturnering (uten maskiner) i 2018 og 2019 (*Geopolitical Forecasting Tournament*). Til forskjell fra GJP, der deltagerne bare fikk betalt for å delta, konkurrerte deltagerne i disse turneringene om pengepremier på en samlet verdi av opptil $250 000.[78] Det er imidlertid ikke publisert noen akademiske studier basert på resultatene fra noen av disse nye turneringene.

Tetlock og Mellers sitt nyeste forskningsprosjekt er GJP 2.0.[79] Prosjektet er en del av et nytt IARPA-program, *FOCUS*, som ser på kontrafaktiske prediksjoner.[80] Kontrafaktiske prediksjoner er påstander om hva som ville skjedd, hvis omstendighetene hadde vært annerledes. Kontrafaktisk prediksjoner danner ofte grunnlaget for erfaringslæring, men det er forsket lite på hvor treffsikre slike kontrafaktisk prediksjoner er og verdien av ulike tilnærminger til erfaringslæring. Et eksempel er de konkurrerende forklaringene på hvorfor Sovjetunionen brøt sammen. Mens konservative vil mene at R. Reagan, som overtok som president i 1981, vant den kalde krigen, vil de liberale hevde at den sovjetiske økonomien holdt på å implodere uansett. Den liberale påstanden hevder dermed implisitt at Sovjetunionen ville kollapset på samme måte, hvis J. Carter hadde blitt gjenvalgt i 1981 og at hans demokratiske etterfølger W. Mondale styrt frem til 1989. Prosjektets hypotese er at analytikere kan bli bedre kontrafaktiske forecastere, hvis de blir bedre til å trekke de riktige kausale lærdommene fra fortiden. Dette skal gjøres ved å trene dem på kontrafaktisk prediksjon i simulerte verdener, for så å teste om de også blir bedre i virkeligheten. Det er ikke publisert noen resultater fra GJP 2.0 ennå.

[74] For nettsiden til GJ Open, se https://www.gjopen.com.
[75] Se McHenry, J. (2018), 'Three IARPA forecasting efforts: ICPM, HFC, and the Geopolitical Forecasting Challenge', *Federal Foresight Community of Interest 18th Quarterly Meeting*, 26. jan. 2018.
[76] Se 'Hybrid Forecasting Competition (HFC)', *IARPA*.
[77] Tetlock et al. (2017), 'Bringing probability judgments into policy debates via forecasting tournaments'.
[78] Se 'Geopolitical Forecasting (GF) Challenge', *IARPA*, og 'Geopolitical Forecasting Challenge 2', *IARPA*.
[79] For mer informasjon om GJP 2.0, se https://www.gjp2.org/. For et intervju med Tetlock der han diskuterer FOCUS, se 'Fireside Chat with Philip Tetlock', *Effective Altruism*, 4. feb. 2020.
[80] Se 'Forecasting Counterfactuals in Uncontrolled Settings (FOCUS)', *IARPA*.

6 Implikasjoner

De viktigste funnene fra GJP er at det *er* mulig å forutse politiske hendelser og utviklinger. Som i EPJ fant GJP at noen personer er systematisk bedre til å predikere enn andre, men også at det er mulig å forbedre den aggregerte treffsikkerheten gjennom relativt enkle grep.

Basert på spesielt funnene fra GJP, er det flere tiltak som kan være relevant å implementere i de delene av forsvarssektoren hvor det gjøres forsvars- og sikkerhetspolitiske analyser:

- *Tallfest sannsynlighetsvurderinger* som uansett gjøres, både for å unngå misforståelser om hva som egentlig menes og for å kunne måle hvor godt de egentlig treffer. Dette kan være unaturlig, men det finnes få faglig funderte argumenter for *ikke* å gjøre dette.[81] Selv om tallfesting kan gjøre at vurderinger fremstår med en større grad av sikkerhet enn det er dekning for, er et gjennomgående funn i GJP at probabilistiske resonnering i seg selv bidrar til å øke treffsikkerheten.

- *Identifiser enkeltpersoner med de beste forutsetningene* for å treffe godt, basert på tester av abstrakt resonneringsevne, kognitiv kontroll, tallforståelse, kunnskapsnivå og tenkemåter. Dette kan være ubehagelig, men seleksjon av de best egnede er vanlig i mange andre sammenhenger. Inntil nå har det imidlertid vært vanskelig å skille mellom folk på bakgrunn av prediksjonsevne. Merk dog at det ikke nødvendigvis er de samme personene som er gode til å predikere som er de beste til å lage gode, relevante spørsmål.

- *Lag grupper av personer* som gjør sannsynlighetsvurderingene. La dem helst være anonyme, gi dem verktøy for å kunne dele og diskutere spørsmålene og bruk treffsikkerhet som statusmarkør. For best resultat, lag elitelag med de aller mest treffsikre. Dette øker forskjeller, men var ett av tiltakene som bidro mest til høyere treffsikkerhet. Dette tiltaket trenger ikke kreve mer enn en omorganisering av personer som allerede gjør sannsynlighetsvurderinger som en del av jobben sin.

- *Gi opplæring* i tankefeil som er vanlige i situasjoner med stor usikkerhet og i teknikker for probabilistisk tenkning, som grunnfrekvens, referanseklasser og gjennomsnittet av flere, uavhengige estimater.

- *Vektlegg prediksjoner* fra personer som har truffet best tidligere og som bringer ny informasjon til torgs. Dette er det allerede laget algoritmer for.

Det er imidlertid tre forbehold ved disse anbefalingene som bør studeres nærmere.

Det første forbeholdet er om funnene er overførbare til en norsk forsvars- og sikkerhetspolitisk kontekst. Deltagerne i GJP var stort sett amerikanske, og spørsmålene var laget ut fra et amerikansk etterretningsperspektiv. Det er ikke gitt at de samme individuelle variasjonene vil gjelde

[81] For en gjennomgang og tilbakevisning av vanlige argumenter mot tallfesting av sannsynlighetsvurderinger, se kapittel 2 i Friedman (2019), *War and Chance*.

for en norsk deltagermasse eller norske forsvars- og fagmiljøer. Det er heller ikke gitt at funnene vil være de samme om spørsmålene hadde tatt utgangspunkt i de viktigste aktørene for norsk sikkerhet. Spørsmålene i GJP hadde også et relativt kort tidsperspektiv, på rundt 100 dager i gjennomsnitt, som er mer relevant for etterretning (f.eks. årlige trusselvurderinger) enn forsvarsplanlegging (der forsvarssektorens langtidsplaner normalt har et planperspektiv på fire år).

Hensikten med FFIs prediksjonsturnering er nettopp å etterprøve disse funnene. Her måles treffsikkerheten til det norske forsvars- og fagmiljøet på spørsmål om spesielt krig og konflikt, Russland, USA og økonomi. Tidsperspektivet på spørsmålene er stort sett 6, 12, 24 eller 36 måneder, altså innenfor det EPJ viste at det var mulig å slå tilfeldig gjetning, men betydelig lenger enn det som ble testet i GJP. Ambisjonen i FFIs turnering er begrenset til å måle treffsikkerheten, hvem som treffer bedre enn andre og hva som kjennetegner norske superforecastere, hvis de finnes. Det har ikke blitt gjennomført eksperimenter for å forbedre treffsikkerheten underveis, men det å identifisere hva slags personer som treffer bedre enn andre er en viktig forutsetning for flere av de andre tiltakene, som sette de beste på grupper med hverandre.

Det andre forbeholdet er i hvor stor grad funnene er overførbare fra turneringer til den virkelige verden. Det er ikke gitt at det er samme personer som treffer best i begge situasjoner. Problemet er at i den virkelige verden måles treffsikkerheten svært sjeldent, og det derfor er mulig å fortsette å predikere helt feil uten at det får konsekvenser for senere vurderinger. Turneringer representerer derimot en metode som gjør det mulig måle treffsikkerheten til mange personer samtidig, legge til rette for trening av prediksjonsevne og å identifisere de beste deltagerne. FFI har allerede utviklet verktøy for gjennomføring av prediksjonsturneringer, som med tilpasning kan tas i bruk av andre innenfor sektoren. For beslutningstagere er ikke det nødvendigvis så viktig hvordan prediksjonene er samlet inn, så lenge de treffer relativt sett bedre enn alternativene.

Det siste forbeholdet er at de virkelig store spørsmålene som betyr noe i forvars- og sikkerhetspolitisk sammenheng – som hvilke scenarioer som bør legges til grunn i langtidsplanleggingen – ikke er mulige å måle treffsikkerheten til, fordi krig er et fenomen som skjer veldig sjeldent. Hvis det først skjer, er det også trolig for sent å gjøre noe med antagelsene som lå til grunn. Det er likevel mulig å tenke seg at de samme kognitive evnene, tenkemåtene, kunnskapsnivåene og situasjonelle faktorene som er diskutert her i sammenheng med treffsikkerhet, også henger sammen med *hva* en tror om det hvordan en fremtidig krig og aktørers mest sannsynlige handlemåter vil se ut i forbindelse med forsvarsplanleggingen. Hvis det er tilfellet, kan det være mulig å «predikere prediksjonene» til enkeltpersoner og fagmiljøer som er involvert i slike analyser. Dette vil i så fall kunne gi verdifull innsikt i hvilke tiltak som kan gjøres for å unngå tankefeilen som disse vil være spesielt utsatt for og dermed bidra til å bomme mindre enn nødvendig.

Referanser

‘Edge Master Class 2015: A Short Course in Superforecasting’, *Edge*, 17. aug. 2015–21. sept. 2015. https://www.edge.org/conversation/philip_tetlock-edge-master-class-2015-a-short-course-in-superforecasting-class-i. Besøkt 14. apr. 2021.

‘Fireside Chat with Philip Tetlock’, *Effective Altruism*, 4. feb. 2020. https://www.effectivealtruism.org/articles/fireside-chat-with-philip-tetlock/. Besøkt 14. apr. 2021.

‘Forecasting Counterfactuals in Uncontrolled Settings (FOCUS)’, *IARPA*. https://www.iarpa.gov/index.php/research-programs/focus/focus-baa. Besøkt 14. apr. 2021.

‘Geopolitical Forecasting (GF) Challenge’, *IARPA*. https://www.iarpa.gov/challenges/gfchallenge.html. Besøkt 14. apr. 2021.

‘Geopolitical Forecasting Challenge 2’, *IARPA*. https://www.iarpa.gov/challenges/gfchallenge2.html. Besøkt 14. apr. 2021.

‘Growth mindset’, *Store norske leksikon*. https://snl.no/growth_mindset. Besøkt 14. apr. 2021.

‘How to Be Less Terrible at Predicting the Future’, *Freakonomics*, 14. jan. 2016. http://freakonomics.com/podcast/how-to-be-less-terrible-at-predicting-the-future-a-new-freakonomics-radio-podcast/. Besøkt 14. apr. 2021.

‘Hybrid Forecasting Competition (HFC)’, *IARPA*. https://www.iarpa.gov/index.php/research-programs/hfc?id=661. 14. apr. 2021.

‘Research That Makes You Go Hmmm on…Forecasts and Predictions’, *The Clemmer Group*, 12. jan. 2016. https://www.clemmergroup.com/blog/2016/01/12/research-that-makes-you-go-hmmm-on-forecasts-and-predictions/. Besøkt 14. apr. 2021.

‘The Aggregative Contingent Estimation Program’, *CitizenScience.gov*. https://www.citizenscience.gov/ace-forecasting/. Besøkt 14. apr. 2021.

Atanasov, P., Rescober, P., Stone, E., Servan-Schreiber, E., Mellers, B., Tetlock, P. og Ungar, L. (2013), ‘The Marketcast Method for Aggregating Prediction Market Forecasts’, i Greenberg, A. M., Kennedy, W. G., og Bos, N. D., red., *Social Computing, Behavioral-Cultural Modeling and Prediction* (SBP 2013).

Atanasov, P., Rescober, P., Stone, E., Swift, S. A., Servan-Schreiber, E., Tetlock, P., Ungar, L., og Mellers, B. (2017), ‘Distilling the Wisdom of Crowds: Prediction Markets vs. Prediction Polls’, *Management Science*, 63:3, ss. 587–900.

Atanasov, P., Witkowski, J., Ungar, L., Mellers, B. og Tetlock, P. (2020), 'Small steps to accuracy: Incremental belief updaters are better forecasters', *Organizational Behavior and Human Decision Processes*, 160, ss. 19–35.

Baron, J. Scott, S. Fincher, K. og Metz, S. E. (2015), 'Why does the Cognitive Reflection Test (sometimes) predict utilitarian moral judgment (and other things)?', *Journal of Applied Research in Memory and Cognition*, 4:3, ss. 265–284.

Baron, J., Mellers, B. A., Tetlock, P. E., Stone, E. og Ungar, L. H. (2014), 'Two Reasons to Make Aggregated Probability Forecasts More Extreme', *Decison Analysis*, 11:2, ss. 133–145.

Beadle, A. W. (2021), 'FFIs prediksjonsturnering – datagrunnlag og foreløpige resultater', *FFI-rapport 21/00737* (Kjeller: Forsvarets forskningsinstitutt).

Beadle, A. W. (2021), 'FFIs prediksjonsturnering – spørsmålskatalog', *FFI-rapport 21/00736* (Kjeller: Forsvarets forskningsinstitutt).

Brier, G. W. (1950), 'Verification of Forecasts Expressed in Terms of Probability', *Monthly Weather Review*, 78:1.

Chang, W., Chen, E., Mellers, B. og Tetlock, P. (2016), 'Developing expert political judgment: The impact of training and practice on judgmental accuracy in geopolitical forecasting tournaments', *Judgment and Decision Making*, 11:5, ss. 509–526.

Chen, E., Budescu, D., Lakshmikanth, S., Mellers, B. og Tetlock, P. (2016), 'Validating the Contribution-Weighted Model: Robustness and Cost-Benefit Analyses', *Decision Analysis*, 13:2, ss. 128–152.

Epple, D. og Romano, R. (2011), 'Peer effects in education: A survey of the theory and evidence', *Handbook of Social Economics*, 1, ss. 1053–1163.

Ericsson, K. A., Krampe, R. T. og Tesch-Romer, C. (1993), 'The role of deliberate practice in the acquisition of expert performance', *Psychological Review*, 100:3, ss. 363–406.

Frederick, S. (2005), 'Cognitive Reflection and Decision Making', *Journal of Economic Perspectives*, 19:4, ss. 25–42.

Friedman, J. A. (2019), *War and Chance: Assessing Uncertainty in International Politics* (Oxford University Press).

Friedman, J. A., Baker, J. D., Mellers, B. A., Tetlock, P. E. og Zeckhauser, R. (2018), 'The Value of Precision in Probability Assessment: Evidence from a Large-Scale Geopolitical Forecasting Tournament', *International Studies Quarterly*, 62:2, ss. 410–422.

Haran, U., Ritov, I. og Mellers, B. A. (2013), 'The role of actively open-minded thinking in information acquisition, accuracy, and calibration', *Judgment and Decision Making*, 8:3, ss. 188–201.

Helland-Riise, F, og Martinussen, M. (2017), 'Måleegenskaper ved de norske versjonene av Ravens matriser [Standard Progressive Matrices (SPM)/Coloured Progressive Matrices (CPM)]', *PsykTestBarn*, 2:2.

Horowitz, M., Stewart, B. M., Tingley, D., Bishop, M., Samotin, L. R., Roberts, M., Chang, W., Mellers, B. og Tetlock, P. (2019), 'What Makes Foreign Policy Teams Tick: Explaining Variation in Group Performance at Geopolitical Forecasting', *The Journal of Politics*, 81:4, ss. 1388–1404.

Johansen, I. (2006), 'Scenarioklasser i Forsvarsstudie 2007: En morfologisk analyse av sikkerhetspolitiske utfordringer mot Norge', *FFI-rapport 2006/02664* (Kjeller: Forsvarets forskningsinstitutt).

Kahneman og Tversky (1977), 'Intuitive prediction: Biases and corrective procedures', *Technical Report PTR-1042-77-6* (Virginia: DARPA).

Kahneman, D. (2013), *Tenke, fort og langsomt* (Oslo: Pax Forlag).

Kruglanski, A. W. og Webster, D. M. (1996), 'Motivated closing of the mind: "Seizing" and "freezing."', *Psychological Review*, 103:2, ss. 263–283.

Lipkus, I. M., Samsa, G. og Rimer, B. K. (2001), 'General Performance on a Numeracy Scale among Highly Educated Samples', *Medical Decision Making*, 21:1, ss. 37–44.

McHenry, J. (2018), 'Three IARPA forecasting efforts: ICPM, HFC, and the Geopolitical Forecasting Challenge', *Federal Foresight Community of Interest 18th Quarterly Meeting*, 26. jan. 2018.

Mellers, B., Baker, J., Chen, E., Mandel, D. og Tetlock, P. (2017), 'How generalizable is good judgment? A multi-task, multi-benchmark study', *Judgment and Decision Making*, 12:4, ss. 369–381.

Mellers, B., Stone, E., Atanasov, P., Rohrbaugh, N., Metz, S. E., Ungar, L., Bishop, M. M., Horowitz, M., Merkle, E. og Tetlock, P. (2015a), 'The Psychology of Intelligence Analysis: Drivers of Prediction Accuracy in World Politics', *Journal of Experiment Psychology: Applied*, 21:1, ss. 1106–1115.

Mellers, B., Stone, E., Murray, T., Minster, A., Rohrbaugh, N., Bishop, M., Chen, E., Baker, J., Hou, Y., Horowitz, M., Ungar, L. og Tetlock, P. (2015b), 'Identifying and Cultivating Superforecasters as a Method of Improving Probabilistic Predictions', *Perspectives on Psychological Science*, 10:3, ss. 267–281.

Mellers, B., Tetlock, P. og Arkes, H. R. (2019), 'Forecasting tournaments, epistemic humility and attitude depolarization', *Cognition*, 188, ss. 19–26.

Mellers, Barbara; Tetlock, Philip, og Arkes, Hal R. (2019), 'Forecasting tournaments, epistemic humility and attitude depolarization', *Cognition*, 188, ss. 19–26.

Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., Scott, S. E., Moore, D., Atanasov, P., Swift, S., A., Murray, T., Stone, E. og Tetlock, P. E. (2014), 'Psychological strategies for winning a geopolitical forecasting tournament', *Psychological Science*, 25:4, 1106–1115.

Moore, D. A., Swift, S. A., Minster, A., Mellers, B., Ungar, L., Tetlock, P., Yang, H. H. J. og Tenneye, E. R. (2017), 'Confidence Calibration in a Multiyear Geopolitical Forecasting Competition', *Management Science*, 63:11, ss. 3552–3565.

Mosteller, F. og Youtz, C. (1990), 'Quantifying Probabilstic Expressions', *Statistical Science*, 5:1, ss. 2–12.

Peters, E., Västfjäll, D., Slovic, P., Mertz, C. K., Mazzocco, K. og Dickert, S. (2006), 'Numeracy and Decision Making', *Psychological Science*, 17:5, ss. 407–413.

Reierth, M. G. og Tronstad, J. (2015), 'Effektive team: Kognitiv motivasjon og maksimerings betydning for hvordan team arbeider og presterer' (Norges handelshøyskole).

Satopää, V. A., Baron, J., Foster, D. P., Mellers, B. A., Tetlock, P. E. og Ungar, L. H. (2014b), 'Combining multiple probability predictions using a simple logit model', *International Journal of Forecasting*, 30:2, 344–356.

Satopää, V. A., Jensen, S. T., Mellers, B. A., Tetlock, P. E. og Ungar, L. H. (2014a), 'Probability aggregation in time-series: Dynamic hierarchical modeling of sparse expert beliefs', *The Annals of Applied Statistics*, 8:2, ss. 1256–1280.

Satopää, V., Pemantle, R. og Ungar, L. (2015), 'Modeling Probability Forecasts via Information Diversity', *Journal of the American Statistical Association*, 111:516, ss. 1623–1633.

Shipley, W. C., Gruber, C. P., Martin, T. A. og Klein, A. M. (2009), *Shipley-2 Manual* (Western Psychological Services).

Surowiecki, J. (2005), *The Wisdom of Crowds* (NY: Anchor Books).

Tetlock, P. (2005), *Expert Political Judgment: How Good Is It? How Can We Know?* (Princeton: Princeton University Press).

Tetlock, P. E. (1998), 'Close-call counterfactuals and belief-system defenses: I was not almost wrong but I was almost right', *Journal of Personality and Social Psychology*, 75:3, ss. 639–652.

Tetlock, P. E. (2010), 'Second Thoughts about Expert Political Judgment: Reply to the Symposium', *Critical Review*, 22: 4, ss. 467–488.

Tetlock, P. E. (2017), *Expert Political Judgment: How Good Is It? How Can We Know?* (New Jersey: Princeton University Press).

Tetlock, P. E., Mellers, B. A. og Scobilic, J. P. (2017), 'Bringing probability judgments into policy debates via forecasting tournaments', *Science*, 355:6324, ss. 481–483.

Tetlock, P. og Gardner, D. (2015), *Superforecasting: The Art and Science of Prediction* (London: Random House Books).

Tetlock, P., Mellers, B., Rohrbaugh, N. og Chen, E. (2014), 'Forecasting Tournaments: Tools for Increasing Transparency and Improving the Quality of Debate', *Current Directions in Psychological Science*, 23:4, ss. 290–295.

Webster, D. M. og Kruglanski, A. W. (1994), 'Individual differences in need for cognitive closure', *Journal of Personality and Social Psychology*, 67:6, ss. 1049–1062.

Åtland, K., Beadle, A. W., Diesen, S., Glærum, S., Mørkved, T., Nyhamar, T. og Stenersen, A. (2018), 'Gjennomgang av FFIs scenariogrunnlag for Forsvarets langtidsplanlegging, 2018', *FFI-rapport 18/00669* (Kjeller: FFI). (BEGRENSET).

About FFI

The Norwegian Defence Research Establishment (FFI) was founded 11th of April 1946. It is organised as an administrative agency subordinate to the Ministry of Defence.

FFI's MISSION

FFI is the prime institution responsible for defence related research in Norway. Its principal mission is to carry out research and development to meet the requirements of the Armed Forces. FFI has the role of chief adviser to the political and military leadership. In particular, the institute shall focus on aspects of the development in science and technology that can influence our security policy or defence planning.

FFI's VISION

FFI turns knowledge and ideas into an efficient defence.

FFI's CHARACTERISTICS

Creative, daring, broad-minded and responsible.

Om FFI

Forsvarets forskningsinstitutt ble etablert 11. april 1946. Instituttet er organisert som et forvaltningsorgan med særskilte fullmakter underlagt Forsvarsdepartementet.

FFIs FORMÅL

Forsvarets forskningsinstitutt er Forsvarets sentrale forskningsinstitusjon og har som formål å drive forskning og utvikling for Forsvarets behov. Videre er FFI rådgiver overfor Forsvarets strategiske ledelse. Spesielt skal instituttet følge opp trekk ved vitenskapelig og militærteknisk utvikling som kan påvirke forutsetningene for sikkerhetspolitikken eller forsvarsplanleggingen.

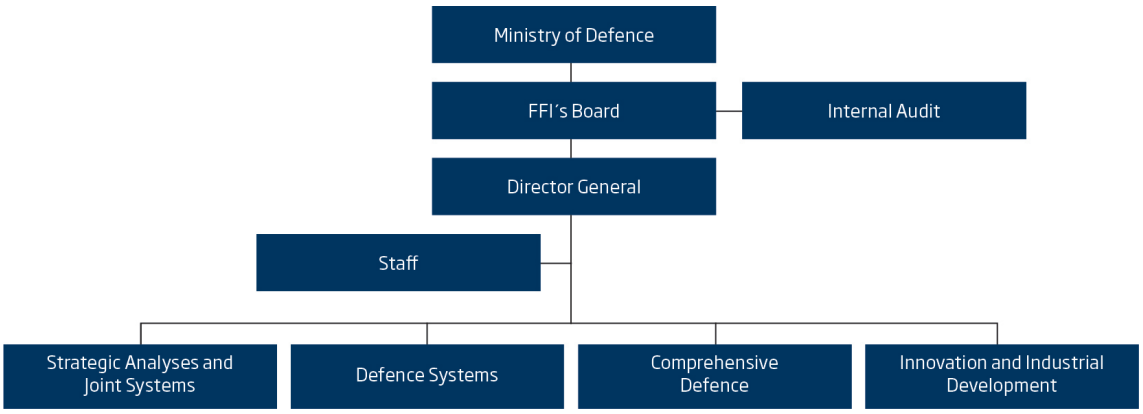
FFIs VISJON

FFI gjør kunnskap og ideer til et effektivt forsvar.

FFIs VERDIER

Skapende, drivende, vidsynt og ansvarlig.

FFI's organisation



Forsvarets forskningsinstitutt
Postboks 25
2027 Kjeller

Besøksadresse:
Instituttveien 20
2007 Kjeller

Telefon: 63 80 70 00
Telefaks: 63 80 71 15
Epost: ffi@ffi.no

Norwegian Defence Research Establishment (FFI)
P.O. Box 25
NO-2027 Kjeller

Office address:
Instituttveien 20
N-2007 Kjeller

Telephone: +47 63 80 70 00
Telefax: +47 63 80 71 15
Email: ffi@ffi.no

FFI Forsvarets forskningsinstitutt