OFFICE OF THE DIRECTOR OF NATIONAL INTELLIGENCE

Forecasting Counterfactuals in Uncontrolled Settings (FOCUS)
Proposers' Day

Paul E. Lehner, Ph.D.

October 19, 2017

Disclaimers

This presentation is provided solely for information and planning purposes

The Proposers' Day does not constitute a formal solicitation for proposals or proposal abstracts

Nothing said at Proposers' Day changes the requirements set forth in a BAA

A BAA supersedes anything presented or said by IARPA at the Proposers' Day

Goals

Familiarize participants with IARPA's interest in the FOCUS program.

Please ask questions & provide feedback, this is your chance to alter the course of events.

Foster discussion of complementary capabilities among potential program participants, AKA teaming. Take a chance, someone might have a missing piece of your puzzle.

Questions

During this session, questions should be recorded on note cards. They will be answered for everyone’s benefit at a later point in the presentation.

If/when a BAA is released, questions can only be submitted to the email address provided in the BAA and will only be answered in writing on the program website.

Agenda

Time	Topic	Speaker
9:00am – 9:30am	Registration and Check In	
9:30am – 9:45am	IARPA Overview and Remarks	Dr. Paul Lehner Chief of T&E, IARPA
9:45am – 10:45am	FOCUS Program Overview	Dr. Paul Lehner Program Manager, IARPA
10:45am – 11:15am	Break	
11:15am – 11:45am	Doing Business with IARPA	Acquisition Team
11:45am – 12:15pm	FOCUS Program Questions & Answers	Dr. Paul Lehner Program Manager, IARPA
12:15pm – 1:15pm	No-Host Lunch	
1:15pm – 3:00pm	Poster Session, Networking, and Teaming Discussions	Attendees (**No Government**)

OFFICE OF THE DIRECTOR OF NATIONAL INTELLIGENCE

FOCUS
Program Overview

Paul Lehner, Ph.D.

INTELLIGENCE ADVANCED RESEARCH PROJECTS ACTIVITY (IARPA)

FOCUS Overview

- FOCUS will be a multi-year research and development program.

- FOCUS will develop and empirically evaluate cognitive methods to improve counterfactual forecasting. Counterfactual forecasts are statements about *what would have happened if* different past circumstances had occurred.

- In a lessons-learned context, counterfactual forecasts are usually claims about what *would have worked better* in past circumstances. Conclusions about what should have been done before, become lessons about what should be done next time, which over time may evolve into purported best practices and tradecraft.

- FOCUS will develop and empirically test alternative approaches to structuring the counterfactual forecasting process in ways that can be readily incorporated into lessons-learned activities relevant to improving analyses and analytic tradecraft in complex domains such as geopolitical analysis.

- Research results should be broadly applicable to any discipline or organization that routinely engages in formal lessons learned activities.

Background and Motivation

Counterfactuals and learning lessons from experience

- Consider the typical elements of a post-mortem/red-team/lessons-learned/what-went-wrong analysis – a lessons-learned analysis (LLA)

– Identify outcome(s) of interest	(e.g. Intelligence failure)	
– Assess why outcome occurred	(e.g. Failure to check assumptions)	Causal inference
– *Assess what would have worked*	*(e.g. Assumption checking)*	*Counterfactual forecast*
– General prescription	(e.g. More assumption checking)	Cause-effect forecasts

- Counterfactual reasoning in a specific instance (what should have been done last time) is the basis for general prescriptions about what to do in the future.

- General prescriptions often evolve to become “best practices”, “standards of practice”, “standards of care”, “tradecraft”, etc.

- Unfortunately ….

Very often we learn the wrong lessons

- In **medicine**: "…*Of the 363 articles testing standard of care, 146 (40.2%) reversed that practice, whereas 138 (38.0%) reaffirmed it.*" (Prasada, … *A Decade of Reversal…* Mayo Clin Proc. 2013;88(8):790-798.)

- In **psychotherapy**, verbal therapy is pretty effective, but success largely uncorrelated with experience, education level, school of therapist or just about anything else practitioner experience tells them is important (Wampold, *The Great Psychotherapy Debate*, 2001)

- In **law enforcement,** experienced interrogators are no more accurate in detecting lies than untrained college students. Also true in detecting false confessions – although everyone better if video is turned off. Behavioral cues of deception are in fact uncorrelated with deception (Fein, et.al. *Educing Information*, 2006 for review).

- In **geopolitical analysis**, forecasts by domain experts are no more accurate than non-expert forecasts (Tetlock, *Expert Political Judgment*, 2005)

- In **management**, consensus forecasts from meetings of experts are less accurate than taking average (Armstrong, How to make better forecasts …, *Intl Jrnl Applied Forecasting*, 2006))

Pick your discipline. Best practices, acquired through years of carefully ***interpreted experiences***, usually turn out to be measurably ineffective or counterproductive

Summary of previous empirical research on how to learn the right lessons.

- Could not find any empirical research on which approaches to counterfactual reasoning yield more accurate counterfactuals.

- Could not find any empirical research on which lessons-learned methods yield more accurate lessons
  - Many methods, papers, and lessons-learned centers
  - Derive their lessons-learned best practices from lessons learned from experiences

- Given general results on best practices, it is reasonable to conclude that we have almost certainly learned the wrong lessons about how to learn the right lessons

FOCUS will begin to fill this research gap; and hopefully establish a research paradigm that will continue to be used for future research

Relating Counterfactuals and Causality

- There is no accepted definition of the word "cause"
- In philosophy most "theories of causality" (aka alternative definitions) involve a mix of two separate concepts – probability and counterfactuals

A before B then A causes B if …	Deterministic	Probabilistic
Non-counterfactual	If A is True then B will always occur	If A is True then B is more likely to occur
Counterfactual (What would have happened if…)	Both A and B occurred, but if A had not occurred, then B would not have occurred	Both A and B occurred, but if A had not occurred then B would have been less likely to occur

Disclaimer: Over-simplification, many nuances not represented here

- Counterfactual definitions tend to be instance specific, whereas non-counterfactual definitions often apply across instances

Causality and counterfactuals are so interwoven that improvements in accuracy of causal and counterfactual reasoning should be coordinate

Pragmatically, the word “cause” is used differently in different disciplines

- In medicine – “cause and effect” relationships are defined probabilistically (across instances) where randomized controlled trials (RCT) can be used to determine that an effect is more likely to occur if the cause is present than if it isn’t (non-counterfactual probabilistic).

- In electrical/mechanical/… engineering – “cause and effect” is very deterministic – removing the cause (in a specific instance) removes the effect (counterfactual deterministic)

- In policy research, statistical methods are employed to evaluate whether a public policy is working by trying to estimate the effect size if the policy had not been present (counterfactual probabilistic)

- In historical analysis – “cause” is a common and key construct, with surprisingly little reflection on how it is or should be defined (see M. Hewitson, *History and Causality*, 2014)

- In intelligence analysis – “cause” is a common and key construct, with surprisingly little reflection on how it is or should be defined (IMO)

Example of pragmatic estimate of causal accuracy in analysis

Consider following statement from 2007 NIE on Iraq stability:

> Iraqi society's growing polarization, the persistent weakness of the security forces and the state in general, and all sides' ready recourse to violence
>
> are collectively driving
>
> an increase in communal and insurgent violence and political extremism.

Causes

Causal connection

Effects

- *In general*,
  - Are effects more likely to occur if purported causes are true?
  - If so, by how much?
  - Example Metric – Causal Information Ratio (CIR)
    - Relative chances that effect will occur if cause is true vs. false
    - CIR = 1.0 implies that cause → effect claims not useful

Declassified key judgments in *Prospects for Iraq's Stability: A Challenging Road Ahead* dated January 2007

$$\text{Est-CIR} = \frac{P(E|C)}{P(E|{\sim}C)} = \frac{11/14}{4/8} = \frac{.76}{.5} = 1.5\text{, n.s.}$$

	E	~E
C	11	3
~C	4	4

Causal accuracy in analyses can be empirically assessed by measuring extent to which purported causes correlate with consequences

Estimates of causal accuracy across multiple analyses

Results of IARPA-funded study

When documents were sufficiently clear to extract cause → effect claims

Document title	Publisher	Year	Prediction range	Cause-prediction pairs	Estimated CIR
Section 1- The Global Infectious Disease Threat and Its Implications for the United States	NIE	2000	2000-present	37	P(E|C) = 1/23 P(E|~C) = 0/14 CIR = und
Section 2- The Global Infectious Disease Threat and Its Implications for the United States	NIE	2000	2000-present	42	P(E|C) = 14/42 P(E|~C) = 0/0 CIR = und
Measuring Political Stability in Afghanistan	SMA	2010	2010-present	7	P(E|C) = 3/3 P(E|~C) = 0/4 CIR = und
2009 STRATFOR US Jihadist War	STRATFOR	2008	2009	11	P(E|C) = 3/6 P(E|~C) = 3/5 CIR = 0.833
2011 STRATFOR Middle East & South Asia	STRATFOR	2010	2011	11	P(E|C) = 0/5 P(E|~C) = 5/6 CIR = 0.00
2008 STRATFOR South Asia	STRATFOR	2007	2008	16	P(E|C) = 12/15 P(E|~C) = 1/0 CIR = 0.933
2002 Annual Forecast: New Priorities Reshuffle the Global Deck	STRATFOR	2001	2002	17	P(E|C) = 6/10 P(E|~C) = 3/7 CIR = 1.400
Further JID forecasts for 2006	Jane's	2006	2006	18	P(E|C) = 8/12 P(E|~C) = 2/6 CIR = 2.00
US and Iran roadmap to conflict	Jane's	2007	2007-2009	11	P(E|C) = 5/8 P(E|~C) = 2/3 CIR = 0.938

	E	~E
C	52	72
~C	16	29

P(E|C) = 52/121 = .420

P(E|~C) = 16/45 = .356

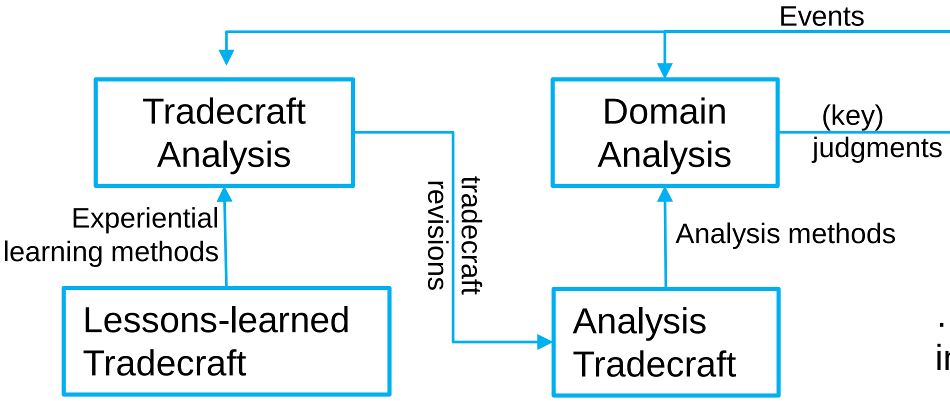
CIR = .420/.356 = 1.18

Overall CIR = 1.18

"CIR = 1.0 implies that cause → effect claims not useful"

Causal and counterfactual accuracy are coordinate – and there is plenty of room to substantially improve both

FOCUSing on Intelligence Analysis



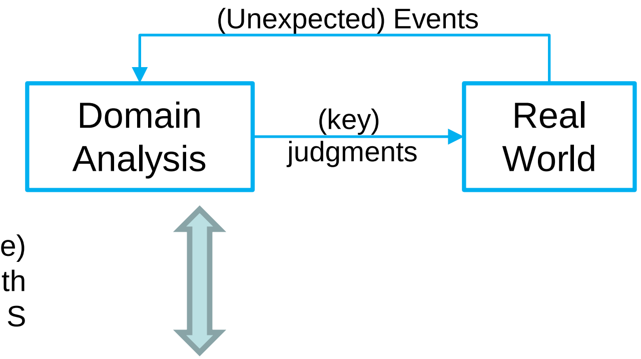
Analysts continuously update their domain understanding and judgments …

… by applying formal and informal analytic tradecraft …

… that evolves through individual and community experiences …

… usually based on a thoughtful and systematic review of those experiences.

Counterfactual reasoning in domain analysis

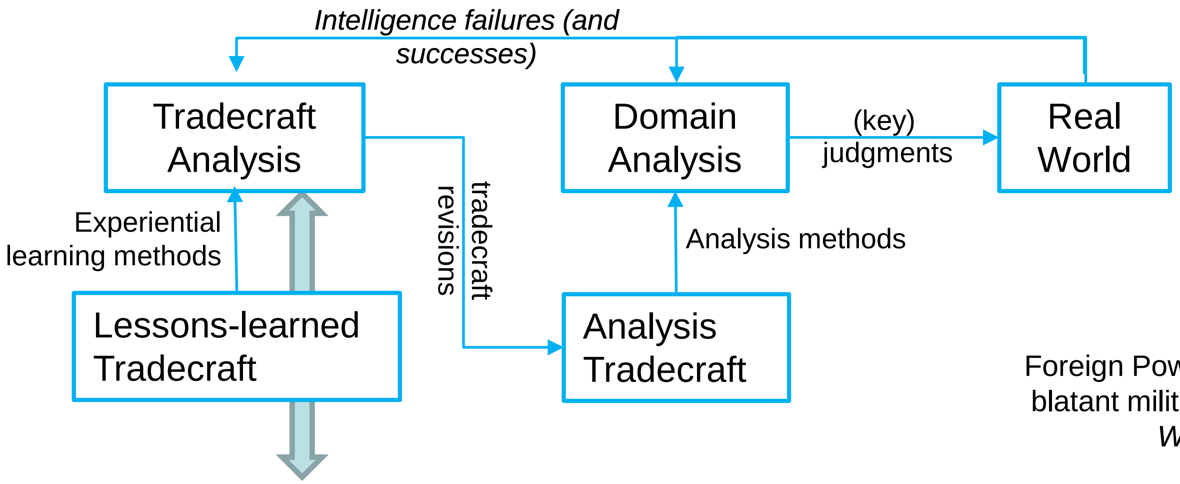


(Made up Example)
Foreign Power R surprised analysts with
blatant military intervention in Country S

R leadership had strong interest in demonstrating regional influence
Should have place more importance of regional influence to R
R's interest in region and showing influence will drive decisions
R will readily consider military intervention elsewhere in region

causal inference (why events happened)
counterfactual forecast (what would have worked)
cause-effect forecasts (key drivers)
event forecasts (what will happen)

Counterfactual reasoning in tradecraft analysis



Foreign Power R surprised analysts with blatant military intervention in Country S
Why were analysts surprised?

causal inference (why failure happened)	Mistakenly assumed avoiding sanctions more important than regional influence
counterfactual forecast (what would have worked)	*Assumption checking would have caught mistake*
cause-effect forecasts (tradecraft benefits)	Formal assumption checking is critical to discovering bad assumptions
event forecasts (accuracy claims)	Apply Key Assumption Checks routinely to reduce chances of similar failures

In the long run, accurate counterfactual reasoning is critical to improving to both analyses and analytic tradecraft

FOCUS Research Objectives

- Develop counterfactual forecasting processes that are …
- Composed of individual component cognitive support methods
- Specifically applicable to improving both domain analysis and tradecraft analysis
- Generally applicable a to diversity of lessons-learned problems
- Yield mostly accurate causal and counterfactual conclusions in diverse domains

Some types of component methods

- Methods that *reduce cognitive biases* related to the attribution of causality, such as for example *fundamental attribution error* where human error is quickly blamed without adequate consideration of other possible causes;

- Methods that adapt forecasting methods to counterfactual forecasting, such as for example adapting *crowd wisdom* or *structured analogy* methods to counterfactual forecasting

- *Structured argumentation* applied to counterfactual reasoning

- *Brainstorming* methods to generate a wide spectrum of possible causal explanations or a wide diversity of possible counterfactual forecasts

- Methods inspired by *counterfactual logics* and the nearest possible worlds semantics that is characteristic of such logics

- Methods that *reduce memory biases*, such as structuring the questions in an interview to minimize hindsight bias of analysts who may inaccurately recall that they anticipated events (e.g. an analysis failure) that they did not in fact anticipate

- Methods adapted from *historical research* engaged in construction of counterfactual histories

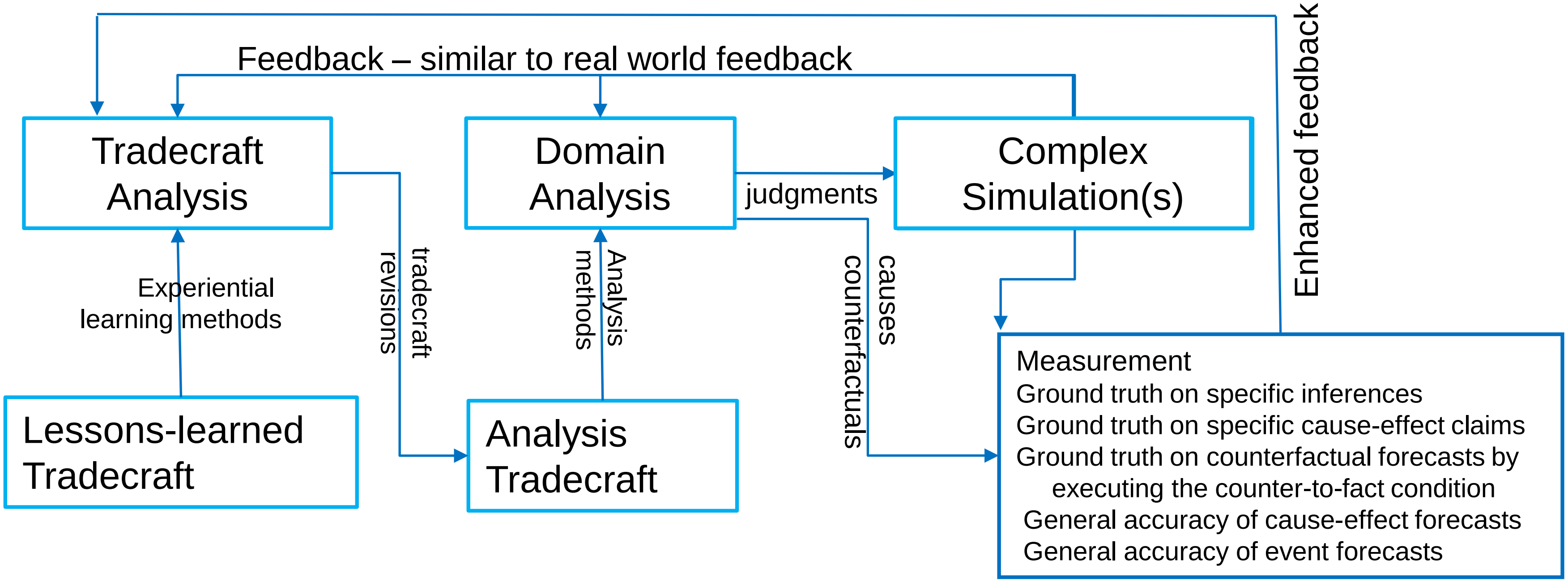
- Methods currently in use in (best practice) lessons-learned processes

Illustrative Example of Counterfactual Process

- Example Question: Would R have invaded S if R did not have a strong interest in establishing regional influence?

- Possible sequence of component methods:
  - Method 1: Enumerate list of possible causes and counter causes (brainstorming)
    - Generate list of factors leading R to invade S, other than regional influence
    - Generate list of factors leading R to not invade S
  - Method 2: Review causes for attribution bias (cognitive bias reduction)
    - Review whether listed causes focus too much on attributional factors, if so
    - Generate additional non-attributional factors (e.g. economic benefits)
  - Method 3: Generate multiple counterfactual scenarios (using structured arguments)
    - Enumerate possible (past) futures based on the various factors
  - Method 4: Review similar historical cases (most similar world semantics, historical analysis methods)
    - Cases where leadership wanted to establish influence but did not invade
    - Cases where leadership did not want to establish influence but still invaded
  - Method 5: Crowd wisdom assignment of past future probabilities (forecasting method)
    - Individuals assign and then average the individual forecast probabilities to generate a scenario forecast
  - Average of probabilities of past futures w/o influence = probability would have invaded anyhow.

Evaluating Counterfactual (and Causal) Accuracy



Measurement
Ground truth on specific inferences
Ground truth on specific cause-effect claims
Ground truth on counterfactual forecasts by executing the counter-to-fact condition
General accuracy of cause-effect forecasts
General accuracy of event forecasts

Ques. How do we measure the accuracy of real world counterfactuals?
Ans. We don’t. But that is not our problem. Rather …

Objective is to measure *overall accuracy* of counterfactual reasoning *methods*. Across multiple domains.

And for that we don’t need (or want) the real world.

Features of the Simulations

- Realistically complex and nuanced
  - Tests/experiments should be as challenging to analysts as the real world
  - Causality in simulation will be substantially deeper than the level of analysis
- Unrealistic
  - Tradecraft/reasoning methods to be tested are supposed to be *general* methods that help analysts to reason about whatever problem or domain they interact with
  - To test general applicability, the methods should be tested against a variety of simulated worlds
  - Some simulations may look like the real world, but have non obvious, nuanced differences (e.g. certain geopolitical actors are really good people who are just misunderstood).
- Maximally re-playable under varying conditions
  - Remove/modify specific purported causes and replay events
  - Insert new actions and replay events
  - Hi frequency replays to determine validity of probabilistic causal and counterfactual claims
- Multiple simulated domains (e.g. geopolitical, city building, fantasy world)
  - Each simulation will create a rich (over simulated time) history, with alternative parameterization for restarting history

Research Phases

Phase 1: ***Develop and refine counterfactual reasoning methods*** **(18 Months)**

- Performer analyst teams develop and refine their counterfactual reasoning methods while performing analyses on complex domain simulations (CDS). Small teams work the analysis tasks, obtain (enhanced) feedback, refine their methods.
- Decision points:
  - Month 9: Review initial results and determine whether performer is making any progress
  - Month 18: Select subset of performers with the most promising methods
- Output: Well-defined counterfactual reasoning methods

Phase 2: ***Controlled confirmatory experiments to estimate impact*** **(9 Months)**

- Controlled experiments with a mix of performer and T&E team participants
- Using CDS, T&E team will have ground truth on counterfactual forecasts, individual key drivers and event forecasts
- Output: Empirical measurement of accuracy and impact of improved counterfactual reasoning

Phase 3: ***Estimating impact on real world forecasting and causal analysis*** **(9 Months)**

- Real world forecasting challenge where participants generate event and causal forecasts, obtain feedback, and apply counterfactual reasoning methods to update their analysis and methods.
- T&E team measures forecast accuracy and estimates aggregate accuracy of cause-effect claims.

Phase 1: Method Development

- Objective: Develop and refine, through experiential learning in simulated domains, counterfactual and causal reasoning methods that work reliably well.

- *"through experiential learning"?!?* – after everything I said previously!

- Recall that the problems with experiential learning described earlier relate to *interpreted experiences* – when it's a matter of judgment and interpretation as to whether an outcome was a success or failure and why.

- Early in Phase 1, performers will receive unambiguous feedback on their causal and counterfactual conclusions – and should be able to quickly refine their causal and counterfactual reasoning methods.

- Once we have counterfactual reasoning/experiential learning methods that do work, *then* we can apply those methods to contexts requiring interpreted experiences. Success in those contexts will be measured in Phases 2 and 3.

Phase 1 Analysis Tasks

- Performer prescribes methods to
  - Generate (probabilistic) counterfactual forecasts (both domain and tradecraft)
  - Generate/update list of key causal domain drivers
  - (Possibly) address other analytic tasks (e.g. hypothesis evaluation, forecasting)

- Iterate through a series of simulated time periods, at each iteration
  - Review reports, historical data, previous analyses, relevant data on other simulated regions, … this will be a rich simulated history,
  - Update prescribed methods
  - Apply prescribed counterfactual forecasting method to answer T&E "what would have happened if …." questions
  - Update list of key domain causal drivers
  - Apply prescribed counterfactual forecasting method to answer T&E "what would you have concluded instead …." tradecraft questions
  - Address current analysis task
  - Receive enhanced feedback on the accuracy of counterfactual forecasts and key causal drivers

- Performers will apply and refine their methods as they perform analyses and iterate through history on multiple different simulated histories in multiple simulated domains.

Examples of Phase 1 enhanced feedback

- Assume that the R invaded S surprise example occurred in a geopolitical simulation, then the simulation could …
- Set counterfactual as true and replay history perhaps multiple times (e.g. remove ‘strong interest’ from R leadership and replay history)
  - Provide direct accuracy feedback on counterfactual forecasts, supporting rapid improvement to counterfactual reasoning processes
- Provide unrealistically detailed information (e.g. inside information leaks discussing true motivations of R leadership)
  - Provides direct feedback on contributing factors and causes
- Provide feedback on specific cause-effect relationships
  - Provides direct feedback on purported cause-effect claims that drive counterfactual forecasts
- Early in Phase 1 feedback will be extensive – we expect you to learn from experience, but we remove the need to ***interpret*** that experience
- As Phase 1 progresses, extended feedback will lesson, with last few rounds *not* including any extended feedback – by then your methods should be much better at learning from interpreted experiences

Phase 2: Rigorous Confirmatory Experiments

- Objective: Experimentally test whether the methods developed in Phase 1 yield accurate causal and counterfactual conclusions in settings with levels of feedback characteristic of real world analysis.

- Performer prescribed methods will be applied to analyses in simulated worlds like in Phase 1 except
  - In the contexts of controlled experiments where prescribed methods will not be updated during each experimental session
  - Participants/analysts will be provided by both Performers and Government team
  - Many of the experiments will be designed to address counterfactual reasoning and lessons learned in tradecraft analysis, where as Phase I emphasized domain analyses

- Emphasis in Phase 2 is using simulations to rigorously measuring accuracy of selected methods … whereas Phase I used the same simulations to provide performers with feedback to help them refine their methods.

Metrics (Phases 1 and 2)

- Accuracy of domain counterfactual forecasts (primary)
  - Probabilistic forecasts measured against (probabilistic) ground truth
  - Ground truth probabilities measured via multiple simulation runs with well-understood (by performers) parameters

- Accuracy of list of key causal drivers (secondary)
  - Proportion of listed causal drivers that are true
  - Ground truth determined via multiple simulation runs with well-understood parameters (key driver: small change in input → big change in output).

- Accuracy of tradecraft counterfactual forecasts (secondary)
  - Estimated a-periodically during Phase 1 by Government team
  - Measured as part of experimental design during Phase 2

Operationally defining any fuzzy concepts

- As described in previous slide, concepts of key causal driver and counterfactual ground truth are admittedly fuzzy.

- FOCUS will operationally define these (and other fuzzy concepts) by how they are measured/determined in each simulation, for example
  - “Key cause” if small (<x%) change in input value → large (>y%) change in output, under conditions of …
  - “Removing strong interest” defined by changing parameters such that …

- Working with performers, we will iterate the operational definitions to ensure meeting the following criteria:
  - Clearly understood by performers
  - Clearly measurable in testing
  - Conform to a general intuitive understanding of these concepts in practice

Scoring (Phases 1 and 2) and Objectives

- Let
  - P be ground truth counterfactual/causal probability [.8]
  - p be the forecast [.4]
- Then
  - Precision: Proportion of forecast obtained = If(pi=0,1,min(1,Pi/pi)) [1]
  - Recall: Proportion of ground truth forecasts = If(P_i=0,1,min(1,p_i/P_i)) [.4/.8 = .5]
  - F1: Aggregate of Precision and Recall = 2*(Pr*Re)/(Pr+Re) [1/1.5 = .67]
- Relation to standard probabilistic scoring rules
  - Not a proper scoring rule
  - Score is a function of the ratio of the true and forecast probability
  - Whereas standard rule, such as squared error, is a function of the difference
- Reflects view that proportion of ground truth forecasted is qualitatively correct measure
- Open to suggestions

Technical Objectives/Milestones

	Q2	Q3	Q4	Q5	Q6
Counterfactual forecasts (F1)	.65	.75	.8	.85	.9
Key causal drivers (F1)	.6	.65	.7	.75	.8

Phase 3: Back to the Real World

- Objective: Determine the extent to which improved counterfactual forecasting yields better domain analysis and tradecraft analysis on real world analysis problems.

- Analysts will engage in a sequence of analyses in a few real world domains where the analysis tasks will include
  - Event forecasting
  - Generating and updating key causal drivers
- and where counterfactual forecasting/lessons-learned method developed in Phase 1 and 2 will be employed to self-reflect between analyses

- Metrics will be
  - Event forecast accuracy metrics (using a standard metric such as Brier score)
  - Causal Information Ratio (CIR) metric that measures *in general* extent to which purported causes are informative of whether an outcome will occur

- Performance will be measured against a control group or natural control condition (e.g. ongoing analyses).

Summary

- We are looking for a diversity of approaches to improving counterfactual forecasting and lessons-learned analyses

- We anticipate teams will include individuals with expertise/experience in relevant psychological research, domain analysis and lessons-learned analysis

- We expect teams will have a strong plan for working across team members to accomplish the program goals.

- Please do review this brief, the draft BAA and any other materials on the FOCUS website. Please send us your recommendations and suggestions for the BAA – using the prescribed format.

- The BAA will supersede anything presented or said at this Proposers’ Day by IARPA.

Point of Contact

Dr. Paul E. Lehner

Program Manager

IARPA, Office of the Director of National Intelligence

Intelligence Advanced Research Projects Activity

Washington, DC 20511

Phone: (301) 851-7449

Fax: (301) 851-7673

Electronic mail: dni-iarpa-baa-17-08@iarpa.gov

(include IARPA-BAA-17-08 in the Subject Line)

Website: www.iarpa.gov

Questions? Please fill out cards.

Point of Contact

Dr. Paul Lehner

Program Manager

IARPA, Office of the Director of National Intelligence

Intelligence Advanced Research Projects Activity

Washington, DC 20511

Phone: (301) 851-7449

Fax: (301) 851-7672

Electronic mail: dni-iarpa-baa-17-08@iarpa.gov
(include FOCUS-IARPA-BAA-17-08 in the Subject Line)

Website: www.iarpa.gov