REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
06/27/2023	Conference Briefing	

4. TITLE AND SUBTITLE

T&E of Complex AI Use Cases

5a. CONTRACT NUMBER

W56KGU-18-D-0004

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Flo M Reeder; Elena Charnetzki

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

The MITRE Corporation
7515 Colshire Drive
McLean, VA 22102

8. PERFORMING ORGANIZATION REPORT NUMBER

PRS-23-2142

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Office of the Director for Developmental Test, Evaluation, and Assessments

10. SPONSOR/MONITOR'S ACRONYM(S)

OSD DTE&A

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

DISTRIBUTION STATEMENT A. Approved for public release: distribution unlimited.

13. SUPPLEMENTARY NOTES

Submission to ITEA Multi-Domain Operations Workshop

14. ABSTRACT

Complex AI use cases require extra attention. In this talk, we will:
§ Define complex AI
§ Describe how AI T&E for complex AI-enabled systems (AIES) is different from traditional T&E
§ Present a use case which represents a complex AIES
§ Describe the T&E needs of this use case

15. SUBJECT TERMS

Artificial Intelligence, evaluation; testing; complex AI; Artificial Intelligence; AI-enabled systems

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE		15	Susan Carpenito
					19b. TELEPHONE NUMBER *(Include area code)* 781-271-7646

T&E of Complex AI Use Cases

Dr. Florence Reeder

Elena Chartnetzki

June 2023

MITRE | SOLVING PROBLEMS FOR A SAFER WORLD®

Introduction

- Complex AI use cases require extra attention
- In this talk, we will:
  - Define complex AI
  - Describe how AI T&E for complex AI-enabled systems (AIES) is different from traditional T&E
  - Present a use case which represents a complex AIES
  - Describe the T&E needs of this use case

IDB.37/4

بالرفاه البشري وتعزيز الصالح العام المُستَّد في الأهداف الإنمائية للألفية. وترتّب اليونيدو أنشطتها لتعزيز التنمية الصناعية ضمن ثلاث أولويات مواضيعية تتضمّن الحدّ من الفقر من خلال الأنشطة الإنتاجية (المرتبطة بالهدفين 1 و3)، وأنشطة بناء القدرات التجارية (المرتبطة بالهدف 8)، وأنشطة البيئة والطاقة (المرتبطة بالهدف 7).

الهدف 1: القضاء على الفقر المدقع والجوع

8- خلال الفترة من 1990 إلى 2005، انخفض عدد الأشخاص الذين يعيشون بدخل يقلّ عن 1.25 دولار في اليوم من 1.8 إلى 1.4 بليون نسمة. بيد أنه يسترداد وضوحا أن بعض المكاسب الكبرى التي تحقّقت في مكافحة الفقر المدقع ستنحسر نتيجة للانكماش الاقتصادي العالمي. ويُقدَّر أنه في عام 2009، سيزيد عدد الذين يعيشون في حالة فقر مدقع بما يتراوح بين 55 مليون و90 مليون نسمة عمّا كان متوقّعا قبل وقوع الأزمة.

9- ويظلّ الحدّ من الفقر كما يتبين من غايات ومؤشّرات الهدف 1 مسألةً مركزية فيما تقدّمه اليونيدو من دعم إلى البلدان النامية. ويقوم ذلك على فكرة أن السبيل الأكثر فعالية للحدّ من الفقر هو استدامة النمو الاقتصادي الذي يمكن أن يتولّد عن طريق تنمية صناعية يقودها القطاع الخاص. وتمثّل الصناعة جزءاً مهماً من عملية تطوير تنظيم المشاريع والاستثمار التجاري والتقدم التكنولوجي والارتقاء بمستوى المهارات البشرية وإيجاد فرص عمل لائقة. والتنمية الصناعية يمكن أيضاً أن تساعد، من خلال الروابط المشتركة بين القطاعات، على إرساء الأساس لقطاع زراعي أكثر فعالية وكفاءة وقطاع خدمات مزدهر. وتساهم جميع هذه العوامل في إدخال تحسينات على نمو الإنتاجية والنمو لصالح الفقراء مما يفضي إلى تحسين مستويات المعيشة.

10- ويشكّل منظّمو المشاريع، والمنشآت الصغيرة والمتوسطة، المصادرَ الرئيسية للأنشطة الاقتصادية التي تدعم نمو الإنتاجية والحدّ من الفقر في البلدان النامية. وهم لديهم القدرة على توليد أعمال منتجة ولائقة، بالإضافة إلى زيادة الاستثمار داخل الاقتصاد. وقد ظل تطوير قدرات تنظيم المشاريع وتقديم الدعم في مجال السياسات من أجل تطوير القطاع الخاص مكوّناً برنامجياً أساسياً في مكافحة اليونيدو للفقر. ويتحقّق ذلك من خلال تعزيز تنمية الموارد البشرية، ونقل المهارات والمعارف، والربط الشبكي فيما بين منظّمي المشاريع والمنشآت الصغيرة والمتوسطة. وينصبّ الاهتمام أيضاً على صوغ السياسات والبرامج والأطر التنظيمية التي من شأنها أن تفضي إلى بيئة تجارية تؤدي، بدورها، إلى نمو المنشآت الصغيرة والمتوسطة، وهو ما يؤدّي أساساً إلى تحقيق النمو لصالح الفقراء والحدّ من الفقر.

V.10-51455

4

Data sample for example complex use case

MITRE

What is Complex AI

Multiple sources of complexity:

- **Mission context:** operational constraints, environmental conditions, mission-specific risk tolerances, legal and policy considerations
- **Integration with other systems:** multiple AI systems operating in concert, upstream and downstream dependencies with other AI or traditional systems
- **Human-AI interaction:** user requirements for explainability, transparency, interpretability, and confidence values; precision/recall preferences
- **System data:** availability of ground truth data, class imbalances or rare events, presence of sensitive data, model drift

How is Complex AI T&E Different?

- T&E involves multiple layers of testing
  - Multiple AI components or multi-model AI component requires component level, integration, and user-level testing
  - Complex data sources increases the variability of the testing requiring a larger evaluation data set
  - Integration in large systems requires lower degree of error tolerance for subsystems the AI interacts with and more careful attention to integration requirements
  - User-interaction complexities require both human-in-the-loop evaluation and specialized training
- Errors in inputs to or outputs from AI components can cascade
  - Characterizing these is important in the T&E process
- Highest performing AI components may not be most effective in a larger system

MITRE

AI T&E Differences to Consider

- *Risks contextualized in the operational environment:* the possibility of cascading errors, compounding uncertainty, and mismatches between precision/recall preferences when AI components are integrated into an analytic workflow
  - Burden on T&E to provide both success and diagnostic information greater
- *Effectiveness evaluated at all levels:* component-level evaluations, integration-level evaluations, and system-level evaluations for mission driven use cases
  - Requirements from which T&E derives may be more complex and T&E data requirements higher to support both diagnostics and operational success
- *Evaluation throughout the full system lifecycle:* from evaluation considerations prior to development through iterative evaluation after deployment and through sustainment.
  - Early engagement of T&E to influence plans and ensure sufficient resources/SMEs

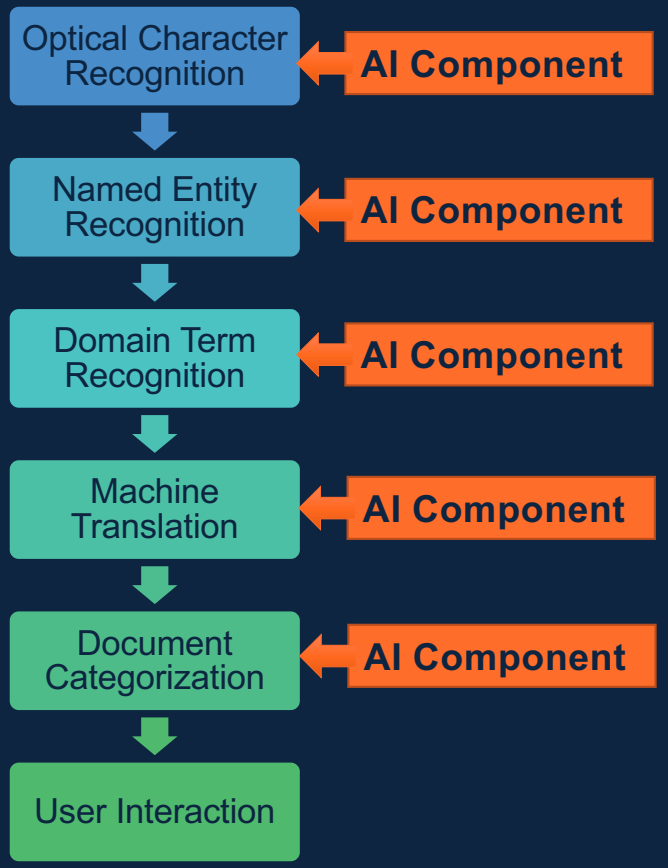
Example MDO Use Case: JADC2 Humanitarian Relief

- Humanitarian assistance and disaster relief operations require close cross-lingual coordination with host nation authorities and Non-Government Organizations (NGOs), often without the benefit of previously established systems
- Crisis operations also generate novel information requirements and require analysts to quickly identify and fuse new sources of information, often in a foreign language
- AI systems can fill critical gaps (e.g., demand for language-enabled analysts with translation and triage skills exceeds availability)
- The outputs of the AIES can be fed into or merged with other information to provide data inputs to mission planning and facilitate a shared operational picture

Why It's Complex

- Multiple AI components working in concert:
  - **Optical character recognition** converts scanned images to text
  - **Named entity recognition** identifies people, places, organizations
  - **Domain term recognition** identifies keywords for a topic
  - **Machine translation** translates from one language to another
  - **Document categorization** bins documents

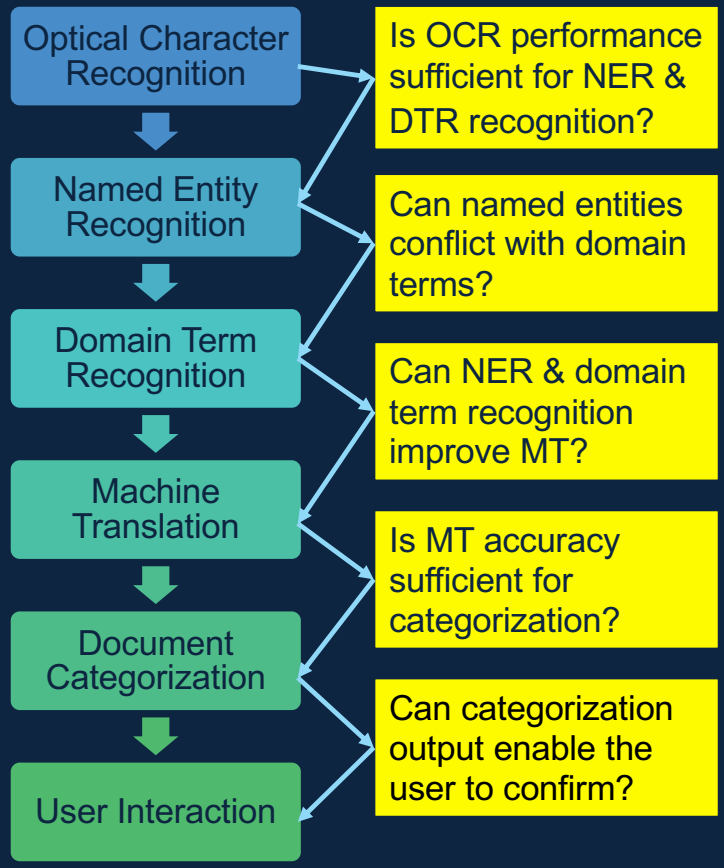
Accuracy for a given component depends on accuracy of previous component



MITRE

T&E for Use Case

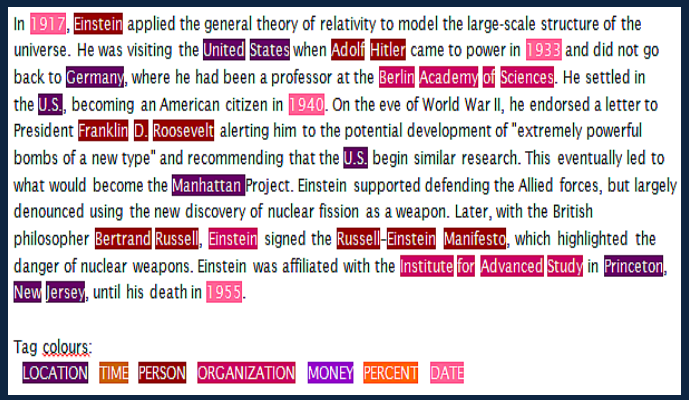
- Each component has a standardized performance metric
  - Optical Character Recognition (OCR) – character or word error rate – data transcribed
  - Named Entity Recognition (NER) – precision/recall on names – data annotated
  - Domain Term Recognition (DTR) – precision/recall on domain terms – data annotated
  - Machine Translation (MT) – BLEU metric for overall document (lacks ground truth).
    - For terms, precision/recall – must be annotated
  - Document categorization – precision/recall per category and precision/recall for triage priority – must be annotated
- Performance metrics, except MT, can help categorize errors effectively



MITRE

T&E for Effectiveness of System

- Did the system prioritize the documents correctly?
  - Ultimate metric is the precision / recall on prioritization decision from the system
  - Precision: percentage of documents in a category that are correct
  - Recall: percentage of documents that belong in a category that are there
  - Precision and recall can be combined into an f-measure and traded off based on user preferences (higher precision comes at the expense of lower recall)
  - A system is most effective when the precision and recall match users' expectations and needs
- Did the system give the analyst sufficient information to confirm the prioritization?
  - Named entities and domain terms often more important
  - User-based testing indicated



MITRE

T&E Issues

- Defining minimal requirements for components is more art than science because system design could either propagate or mitigate errors
  - Highest scoring component model could be less effective, depending on the type of errors – sometimes handled by pre-processing or checking outputs of upstream model prior to downstream one
- AIES component testing increases data curation burden since different annotations are required for each component except MT
  - MT is a technology which has no ground truth (right answer), so evaluation must focus on the requirements of the system which may have ground truth (such as names)
- Analysts and SMEs must identify domain terms and other important information needs from triage
- Analysts must work with developers on user interface to ensure triage result is confirmable – testing involves human-in-the-loop

MITRE

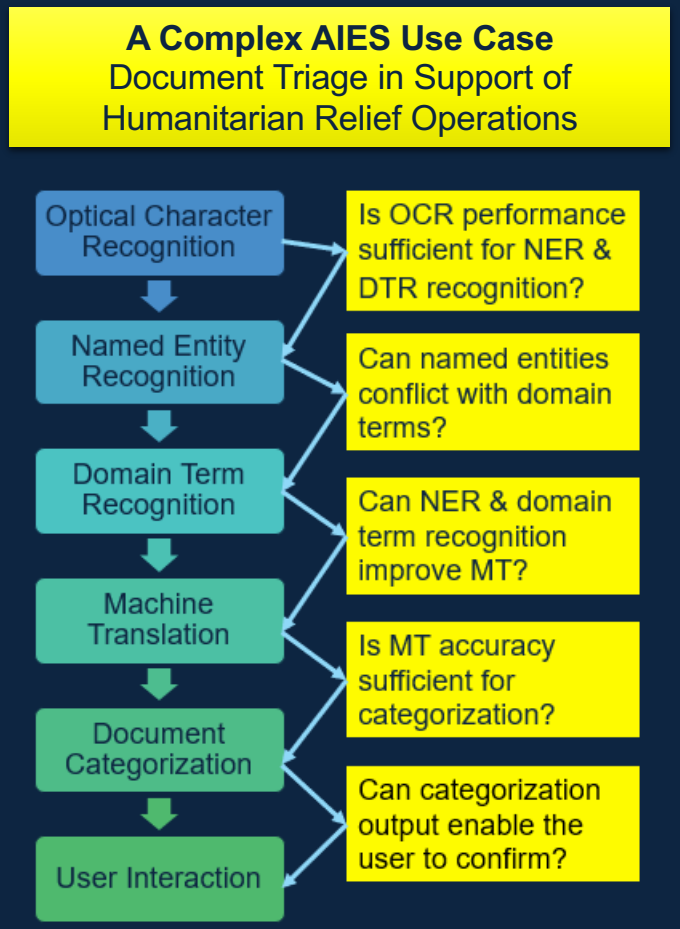
T&E of Complex AI Use Cases: Summary

T&E of AI Enabled Systems must address multiple sources of complexity:

- Multiple AI components operating in concert requires component level, integration, and user-level testing
- Complex data sources increase the variability of the testing requiring a larger evaluation data set
- Integration in large systems requires a higher degree of error tolerance for subsystems AI interacts with
- User-interaction complexities require both human-in-the-loop evaluation and specialized training

Examining a specific use case illuminates the importance of:

- *Risks contextualized in the operational environment:* the possibility of cascading errors, compounding uncertainty, and mismatches between precision/recall preferences when AI components are integrated into an analytic workflow
- *Evaluation throughout the full system lifecycle:* from evaluation considerations prior to development through iterative evaluation after deployment and through sustainment which needs component telemetry for trouble shooting in deployed model.



MITRE

Questions?

MITRE

Backup

MITRE

(ه) أن يعتمد استمارة وثيقة تصميم برنامج أنشطة التنفيذ المشترك (انظر المرفق الثالث)؛

(و) أن يحث الأطراف المُدرجة في المرفق الأول للاتفاقية (أطراف المرفق الأول) على تقديم مساهمات في الصندوق الاستئماني للأنشطة التكميلية بغرض تمويل العمل المتعلق بالتنفيذ المشترك في فترة السنتين 2010-2011، على مستوى يتيح تنفيذ خطة إدارة التنفيذ المشترك لفترة السنتين تنفيذاً كاملاً؛

(ز) أن ينتخب عضواً واحداً وعضواً مناوباً في لجنة الإشراف على التنفيذ المشترك من كل دائرة من الدوائر التالية لفترة سنتين، استناداً إلى الترشيحات الواردة:

'1' الأطراف المُدرجة في المرفق الأول التي تجتاز عملية انتقال إلى اقتصاد السوق؛

'2' الأطراف الأخرى المدرجة في المرفق الأول؛

'3' تحالف الدول الجذرية الصغيرة.

(ح) أن ينتخب عضوين وعضوين مناوبين في لجنة الإشراف على التنفيذ المشترك من الأطراف غير المدرجة في المرفق الأول للاتفاقية لفترة سنتين استناداً إلى الترشيحات الواردة.

11- وحتى تاريخ إعداد هذا التقرير، لم تتمكن لجنة الإشراف على التنفيذ المشترك من تعيين عضو جديد يحل محل العضو الذي استقال من 30 أيلول/سبتمبر 2009 لعدم ورود أي ترشيحات من الفئة المعنية. ولذلك قد يحتاج مؤتمر الأطراف/اجتماع الأطراف إلى انتخاب عضو إضافي من طرف من الأطراف المُدرجة في المرفق الأول يجتاز عملية انتقال إلى اقتصاد السوق ليحل محل العضو المستقيل خلال باقي مدته (انظر الفقرة 45 أدناه).

ثالثاً - العمل المنجز منذ تقرير لجنة الإشراف على التنفيذ المشترك المقدم إلى الدورة الرابعة لمؤتمر الأطراف/اجتماع الأطراف في بروتوكول كيوتو

ألف - موجز للعمل المنجز

12- مع إطلاق إجراء المسار الثاني رسمياً في تشرين الأول/أكتوبر 2006، تحوَّل تركيز لجنة الإشراف على التنفيذ المشترك إلى تنفيذ الإجراء ذاته. وعلى مدى الأعوام الثلاثة الماضية، تناولت اللجنة ورقات المعلومات المتصلة بالمشاريع، كما عملت، بما في ذلك عن طريق فريق الاعتماد التابع لها، على اعتماد الكيانات المستقلة. وعلاوة على ذلك، أصدرت اللجنة، عند اللزوم، إرشادات وتوضيحات تتعلق بكل من إجراء المسار الثاني وإجراء الاعتماد.

MITRE

بالرفاه البشري وتعزيز الصالح العام المُحدَّد في الأهداف الإنمائية للألفية. وترتِّب اليونيدو أنشطتها لتعزيز التنمية الصناعية ضمن ثلاث أولويات مواضيعية تتضمّن الحدّ من الفقر من خلال الأنشطة الإنتاجية (المرتبطة بالهدفين 1 و3)، وأنشطة بناء القدرات التجارية (المرتبطة بالهدف 8)، وأنشطة البيئة والطاقة (المرتبطة بالهدف 7).

الهدف 1: القضاء على الفقر المدقع والجوع

8- خلال الفترة من 1990 إلى 2005، انخفض عدد الأشخاص الذين يعيشون بدخل يقلّ عن 1.25 دولار في اليوم من 1.8 إلى 1.4 بليون نسمة. بيد أنه يزداد وضوحا أن بعض المكاسب الكبرى التي تحقّقت في مكافحة الفقر المدقع ستنحسر نتيجة للانكماش الاقتصادي العالمي. ويُقدَّر أنه في عام 2009، سيزيد عدد الذين يعيشون في حالة فقر مدقع بما يتراوح بين 55 مليون و90 مليون نسمة عمّا كان متوقّعا قبل وقوع الأزمة.

9- ويظلّ الحدّ من الفقر كما يتبين من غايات ومؤشّرات الهدف 1 مسألةً مركزية فيما تقدّمه اليونيدو من دعم إلى البلدان النامية. ويقوم ذلك على فكرة أن السبيل الأكثر فعّالية للحدّ من الفقر هو استدامة النمو الاقتصادي الذي يمكن أن يتولّد عن طريق تنمية صناعية يقودها القطاع الخاص. وتمثّل الصناعة جزءاً مهماً من عملية تطوير تنظيم المشاريع والاستثمار التجاري والتقدم التكنولوجي والارتقاء بمستوى المهارات البشرية وإيجاد فرص عمل لائقة. والتنمية الصناعية يمكن أيضاً أن تساعد، من خلال الروابط المشتركة بين القطاعات، على إرساء الأساس لقطاع زراعي أكثر فعالية وكفاءة وقطاع خدمات مزدهر. وتساهم جميع هذه العوامل في إدخال تحسينات على نمو الإنتاجية والنمو لصالح الفقراء بما يفضي إلى تحسين مستويات المعيشة.

10- ويشكّل منظّمو المشاريع، والمنشآت الصغيرة والمتوسطة، المصادرَ الرئيسية للأنشطة الاقتصادية التي تدعم نمو الإنتاجية والحدّ من الفقر في البلدان النامية. وهم لديهم القدرة على توليد أعمال منتجة ولائقة، بالإضافة إلى زيادة الاستثمار داخل الاقتصاد. وقد ظل تطوير قدرات تنظيم المشاريع وتقديم الدعم في مجال السياسات من أجل تطوير القطاع الخاص مكوّناً برنامجياً أساسياً في مكافحة اليونيدو للفقر. ويتحقّق ذلك من خلال تعزيز تنمية الموارد البشرية، ونقل المهارات والمعارف، والربط الشبكي فيما بين منظّمي المشاريع والمنشآت الصغيرة والمتوسطة. وينصبّ الاهتمام أيضاً على صوغ السياسات والبرامج والأطر التنظيمية التي من شأنها أن تفضي إلى بيئة تجارية تؤدي، بدورها، إلى نمو المنشآت الصغيرة والمتوسطة، وهو ما يؤدّي أساساً إلى تحقيق النمو لصالح الفقراء والحدّ من الفقر.