Joint Activity Testing (JAT)

A New Approach to Human-Machine Testing & Evaluation

Authors: Dane A. Morey & Michael F. Rayo

The Challenges of Human-Machine T&E

As machines continue to behave more like active cognitive agents (e.g., able to make recommendations, interjections, etc.) than passive tools, human-machine systems have become significantly more difficult to evaluate in both laboratory and naturalistic settings. The increasing interconnectedness and interdependence of human-machine systems renders testing and evaluation (T&E) of individual components (i.e., the human or machine alone) increasingly inadequate for understanding joint system performance. For example, machines that provide recommendations can have a strong influence on human cognitive processes and produce behaviors that are both better and substantially worse than if the human operators had been working unaided (Smith et al., 1997). Consequently, these systems are often far more brittle in the real world than their designers intendended, as highlighted by a number of recent high-profile incidents involving highly automated technologies (NTSB, 2017; Eubanks, 2018; Obermeyer et al., 2019; FAA, 2020).

The creeping complexity of these human-machine systems has far outpaced T&E methods, e.g., usability tests (Rayo, 2017). No singular discrete set of observations can capture the full range of difficulties a system will face during operational use. Yet the high costs of running high-fidelity human-machine studies means testing sets will remain inevitably limited. Therefore, the ability to extend T&E insights beyond the limits of testing sets, holistically compare the performance of joint system architectures, and extrapolate trends of brittleness or extensibility becomes critical to mitigate the risks of introducing unintended consequences.

The Joint Activity Testing (JAT) Methodology

We are developing *Joint Activity Testing* (JAT) as a T&E method to directly compare the performance of multiple competing system alternatives, including the performance of individual components (i.e., unaided human, unsupervised AI) as well as the joint human-machine system, in a way that enables insights to extend beyond testing boundaries (Morey et al., 2020). The method compares alternatives across multiple, continuous measures of performance (e.g., efficiency, accuracy, etc.) and challenge to the system (e.g., workload, required tempo of activity, accuracy of available data, etc.) which are built from empirically-based patterns of difficulties in distributed cognitive work (Patterson et al., 2010). Plotting the results of discrete testing cases within this performance-challenge space then facilitates a more confident interpolation, extrapolation, and abstraction beyond the boundaries of the testing set. This analysis can help evaluators anticipate how system performance will change as challenge to the system

increases, holistically compare alternatives across the entire range of potential challenges, assess the sufficiency of the testing set itself, and identify regions of performance and challenge that are likely to yield maximally informative results.

We have now begun to operationalize JAT in multiple intelligence analysis and healthcare settings by pursuing two parallel lines of inquiry: (1) plotting discrete testing cases within the frame of reference performance vs. challenge and (2) modeling performance as challenge increases. Using the results of a recent human-machine teaming study, we showed the feasibility of both aspects of JAT with one dimension of performance and one dimension of challenge (Morey et al., 2020). In several other ongoing projects, we have begun to implement this methodology with multiple dimensions of performance and challenge. The following steps outline the general process of conducting JAT.

1. **Define Performance Measures**. Select meaningful outcomes (often capturing aspects of efficiency, accuracy, or thoroughness) of the joint system that can be measured to assess performance. This becomes the y-axis of each plot.
2. **Define Challenge Measures**. Select meaningful aspects of the domain that can be measured to quantify the degree of challenge the system faces (e.g., Patterson et al., 2010). This becomes the x-axis of each plot.
3. **Identify a Reference**. To contextualize results, it is advantageous to compare the performance of at least two different configurations of the joint system (e.g., with and without machine recommendations).
4. **Select a Testing Set**. A central purpose of JAT is to assess how system performance changes as challenge to the system increases; therefore, the testing set should include a wide range of challenge degrees along the x-axis.
5. **Run the Study**. Ideally, studies should simulate real-world tasks and tools as closely as possible.
6. **Plot the Results**. Each testing case for each participant can be associated with a specific degree of challenge and degree of performance, which can be directly plotted on the graph (figure 1) and used to derive insights.
7. **Fit a Model of Performance** (ongoing research). We continue to explore methods to model how system performance changes as challenge to the system increases.
8. **Compare Model Characteristics** (ongoing research). With a model of performance, it is possible to compare higher-level properties of system performance (e.g., net benefit or slope of decline).

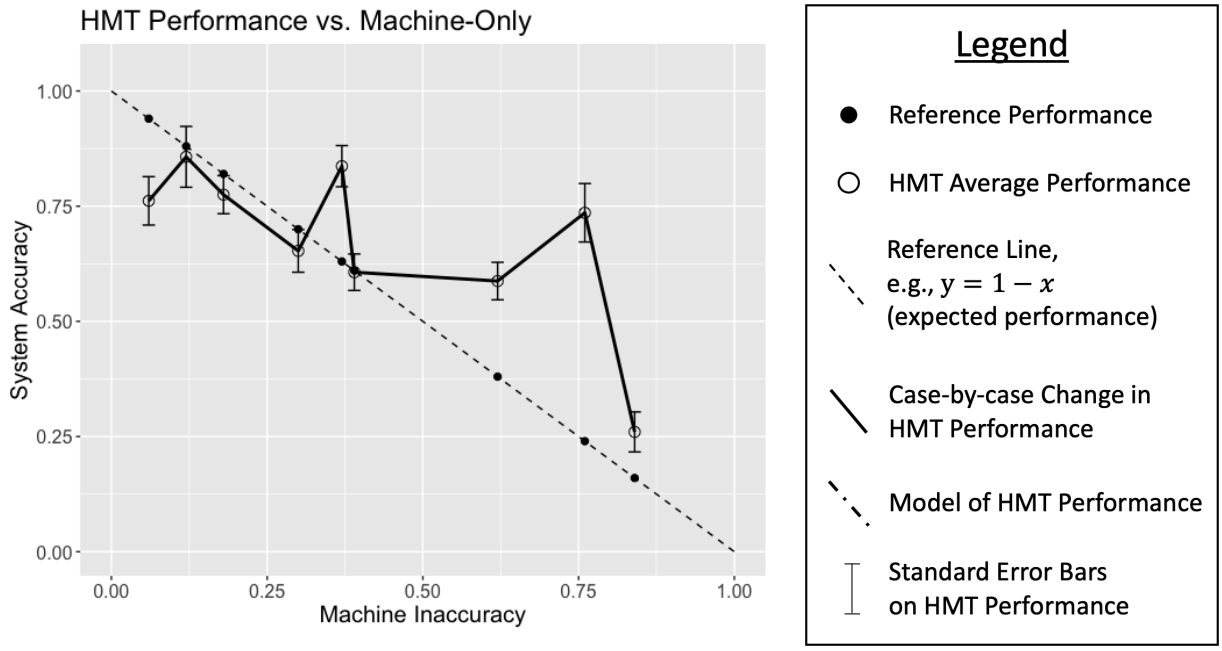

Figure 1. Example graph resulting from JAT (Morey et al., 2020)

Future Directions

Modeling Performance Curves

We are actively exploring ways to fit a model to the quasi-continuous data collected from T&E. This acts as a model of system performance as challenge to the system increases. In addition to the net area (e.g., between the model curves and the reference line), higher-level characteristics like slope may become increasingly informative. We expect two of the high-level parameters of system performance proposed in Morey et al. (2020) to be particularly informative:

- **Net area between two curves** (e.g., model of HMT performance and reference line), calculated as the difference between the integrals of each curve across the entire range of challenge.
- The point at which the slope is maximal and/or the **steepness of the curve**, calculated as the point at which the second derivative is zero and/or the value of the first derivative.

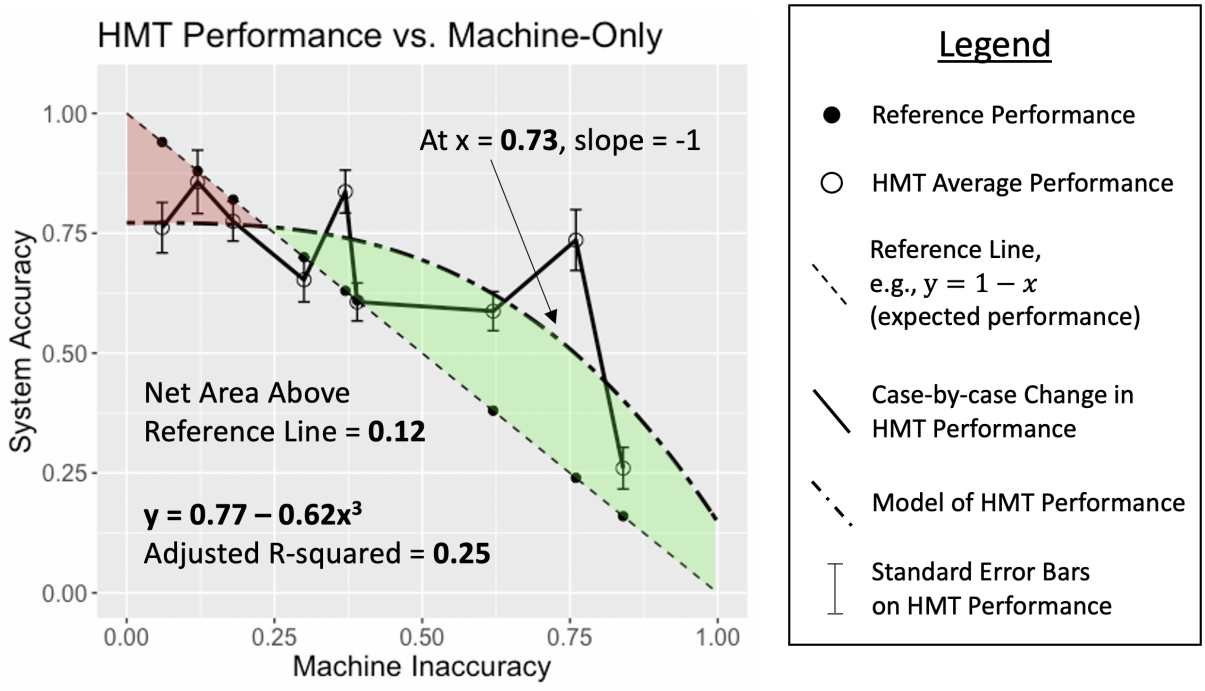

Figure 2. Example graph with performance model

Such a model enables several capabilities crucial for human-machine T&E: (1) interpolating and extrapolating system performance beyond the boundaries of the testing set, (2) holistically comparing alternative system architectures, and (3) assessing the sufficiency of the testing set itself to guide future testing.

Extrapolating System Performance

The model provides a prediction of system performance at any degree of challenge within or outside the range tested in the testing set. While system performance is likely to be noisy, we believe even a crude estimate beyond the boundaries of the testing set is valuable data for evaluators and decision-makers.

Holistically Comparing Alternatives

Calculated characteristics from the model of performance can help evaluators holistically compare competing system architectures. The net area between two curves informs the extent to which one system architecture outperforms another across the entire spectrum of potential challenges. The rate at which performance declines can help inform the degree to which a system exhibits performance that is characteristic of brittleness.

Assessing the Testing Set

Without knowing a priori which regions of challenge are maximally informative, it is important to be able to assess the sufficiency of the testing set to guide future testing. Regions of

performance that appear undersampled or particularly volatile are regions where additional testing would be maximally informative to the model. On the other hand, regions of performance where the curve is relatively flat or performance is more stable may not yield results that significantly reduce the uncertainty in the testing.

Multidimensional JAT

Performance and challenge are not unidimensional; therefore, JAT should include multiple dimensions of performance and challenge. To construct multiple plots, testing sets will need to be constructed so that the differential effects of different challenges can be analyzed. Therefore, the performance curves will likely need to become models of the effects determined by statistical analyses. Figure 3 shows notional examples of JAT in multiple dimensions. We believe that by integrating multiple perspectives of systems, evaluators can better understand to which classes of challenges systems seem to be particularly vulnerable.

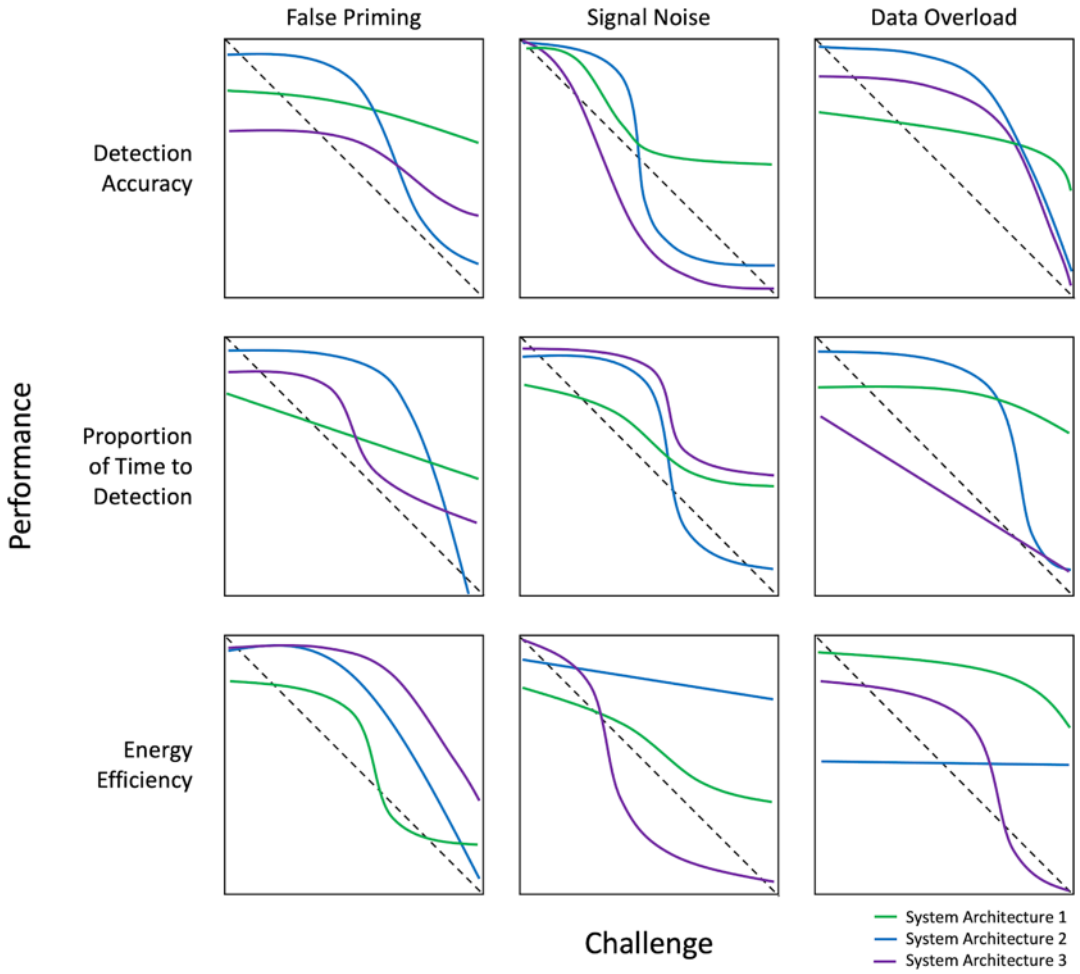

Figure 3. Notional example of multidimensional JAT, adapted from Morey et al. (2020)

References

Eubanks, V. (2018). Automating inequality: How high-tech tools profile, police, and punish the poor. St. Martin's Press.

FAA (2020, August 3). *Preliminary Summary of the FAA's Review of the Boeing 737 MAX*. https://www.faa.gov/news/media/attachments/737-MAX-RTS-Preliminary-Summary-v-1.pdf

NTSB (2017, September 12). *Collision Between a Car Operating With Automated Vehicle Control Systems and a Tractor-Semitrailer Truck Near Williston, Florida, May 7, 2016*. Highway Accident Report. https://www.ntsb.gov/investigations/AccidentReports/Reports/HAR1702.pdf

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453. doi:10.1126/science.aax2342

Patterson, E. S., Roth, E. M., & Woods, D. D. (2010). Facets of complexity in situated work. In J. E. Miller & E. S. Patterson (Eds.). *Macrocognition metrics and scenarios: Design and evaluation for real-world teams*. Ashgate.

Rayo, M. F. (2017, September). Designing for collaborative autonomy: updating user-centered design heuristics and evaluation methods. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 61, No. 1, pp. 1091-1095). Sage CA: Los Angeles, CA: SAGE Publications.

Smith, P. J., McCoy, C. E., & Layton, C. (1997). Brittleness in the design of cooperative problem-solving systems: The effects on user performance. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, *27*(3), 360-371.