An Approach for Visualizing Comparison of Human and AI Decision-Making

Henry Phillips, Alyssa Tanaka, Angela Woods

Soar Technology, Inc

Orlando FL

henry.phillips@soartech.com, alyssa.tanaka@soartech.com, angela.woods@soartech.com

The task of consolidating and quantifying complex decision-making across a multi-attribute space is a challenge (Achkoski et al., 2017; Klein, 2008). This work describes a technique for quantifying human expert multi-faceted decision-making in the context of medical triage into a limited number of dimensions. The analytic and visualization technique described here also serves as a basis for both *machine-readable* and *human-interpretable* comparisons between human and algorithmic decision-making outcomes.

Tactical Combat Casualty Care (TC3) and medical triage scenarios involve complex situations comprised of factors including time pressure, high stakes, and uncertainty (Joint Trauma System, 2020; Klein, Orasanu, Calderwood, & Zsambok, 1993). Medical triage scenarios are an exercise in satisficing, since by its very definition, "triage" refers to the prioritization of limited tasks and resources. In complex triage scenarios, experts will disagree on what set of decisions is truly optimal (Achkoski et al., 2017). This effort used simulated values approximating a dataset of 28 scenario responses from practitioners and laypeople to capture complex triage decisions and translate the factors underlying them to quantifiable, comparable metrics. This modeling is a critical prerequisite for a measurement and visualization tool that could lend itself to comparisons of decisions made by human experts with decisions made by AI systems.

The goal of this effort was design of a simple system for representing and capturing treatment and resource allocation decisions by triage managers and translating those decision data to a limited number of dimensional attribute scores describing the decision-makers. This reduction made it possible to consolidate this representation of the data representing a group of decision-makers in a single multi-dimensional index. Machine learning techniques were then used to evaluate the predictive relationships among scenario characteristics and decision-maker attributes, also discussed (Zheng, Aragam, Ravikumar, & Xing, 2018).

Keywords

Machine learning
Medical triage
Decision-making
Data visualization

ABOUT THE AUTHORS

Dr. Henry L Phillips IV, a Senior Program Manager with SoarTech since 2020, is responsible for identification, management, and alignment of new development opportunities in the areas of AI/ML training tools and capabilities. He joined SoarTech upon retirement as a Navy Commander following a 20-year career as an Aerospace Experimental Psychologist, including service as Executive Officer of the Naval Air Warfare Center Training Systems Division. During his active-duty career, he developed modeling and simulation training tools and capabilities in use by 11,000 warfighters per year, and remains a leader in research and development in human systems, modeling and simulation, training, and intelligent systems. He is a winner of the Chatelier Lifetime Achievement Award, the Team Orlando Collaboration Award for the Squad Overmatch project, the Army HSI Award for his development of the HSI Readiness Level model and was part of the team that won the Boorda Award for Policy for his work on the Live Virtual Constructive Training Fidelity effort. He holds a PhD in Industrial/Organizational Psychology from the University of Houston, and has over 70 peer-reviewed articles, technical reports, and presentations in the areas of training, machine learning, modeling and simulation, item response theory, personnel selection, and psychometrics.

Dr. Alyssa Tanaka, a Lead Scientist with SoarTech since April 2017, conducts artificial intelligence (AI) research for medical domains. She has earned a Ph.D. and M.S. in Modeling and Simulation from the University of Central Florida, Graduate Certificates in Instructional Design and Training Simulations, and a B.S. in Psychology and Cognitive Sciences from the University of Central Florida. She also holds a diploma in Robotic Surgery from the Department of Surgery, University of Nancy, France. At SoarTech, her research focuses on the application of AI for improving medical outcomes. She currently leads a portfolio of DoD funded research efforts focused on developing clinical decision support systems and improving TCCC education and training. Prior to joining SoarTech, Dr. Tanaka led research initiatives directly related to improving the training and education of robotic surgeons through simulation as a Senior Research Scientist at AdventHealth Nicholson Center, the premier location in the world both for training expert practitioners with the daVinci surgical robot and for advancing research in cutting-edge training to target this device.

Angela Woods is a senior software architect with Soar Technology since 2006. She is an expert in systems architecture modeling, learning architecture design, systems integration, distributed systems design, and constraint-based expert models. Her work also extends to data analytics, and temporal data modeling. She holds a bachelor's degree in Computer Science from the University of Washington.

An Approach for Visualizing Comparison of Human and AI Decision-Making

Henry Phillips, Alyssa Tanaka, Angela Woods

Soar Technology, Inc.

Orlando FL

henry.phillips@soartech.com, alyssa.tanaka@soartech.com, angela.woods@soartech.com

Decision-making in medical triage is challenging and complex. In mass casualty situations, medical providers must make critical time-sensitive decisions for which no clear "correct" answer may be obvious (e.g., whom to treat, with what resources, in what order). In such settings, experts may disagree on what the optimal set of decisions for a scenario would look like, or even whether an optimal set of decisions exist. Decisions in such scenarios made by artificial intelligence (AI) algorithms (Andronie et al., 2021) can be assessed by comparing their recommendations and courses of action to the consensus of human decision-makers, *if such comparisons with human expert consensus can be made based on a readily interpretable summary of that consensus*, or lack thereof. The goal of this effort was design of an approach for visualizing human medical triage decision-making and an interpretable mechanism for comparison of solutions to the same scenario proposed by AI algorithmic systems.

Human decision-making in medical triage can be described using multiple theoretical approaches. Tactical Combat Casualty Care (TC3) and medical triage scenarios frequently involve complex situations comprised of factors including time pressure, high stakes, and uncertainty, consistent with the principles of naturalistic decision-making (Klein, 2008; Klein, Orasanu, Calderwood, & Zsambok, 1993). But in cases where those scenarios are complex enough that they have no clear right answers, an even more appropriate approach might be the theory of bounded rationality (Selten, 1999; Simon, 1957), which is based on the foundational assumption that there is never one best course of action, largely because it is impossible to have complete information on any subject. Medical triage scenarios are an exercise in satisficing, since by its very definition, "triage" refers to the prioritization of tasks and resources, when those resources are scarce or insufficient: resources that include medical supplies, available hands, time, and information.

Medical triage is the single most important factor in reducing lives lost during combat casualty management and poses a significant burden on combat medics due to these information gaps, continuously evolving complex challenges and uncertainties in real-world TC3 scenarios. These factors make TC3 scenarios an ideal candidate for the current effort, which focused on capturing these complex decisions and translating the factors underlying them to quantifiable, comparable metrics. TC3 scenario treatment decisions, particularly in complex situations involving multiple casualties, also add the additional complication that many experts will disagree on what the optimal set of decisions for a scenario would look like, or even whether an optimal set of decisions exist. For many of these complex TC3 scenarios, there will be no universally recognized best course(s) of action.

This quantification and modeling is an important prerequisite toward a larger goal of building a measurement and visualization tool that could lend itself to comparisons of decisions made by human experts with decisions made by AI systems. To replicate triage manager decision-making during TC3 situations using AI, a necessary first step is to build a simple, structured, and quantifiable approach to capture and describe the treatment and resource allocation decisions made by triage managers in multiple casualty scenarios. Some work has been done on developing AI to offer triage support, but the point of focus has been on monitoring soldiers' physiology on the battlefield rather than supporting medical *decision makers (DMs)* (Achkoski et al., 2017; Stevanoski et al., 2016). Triage requires rapidly evaluating and making rapid decisions about patient prioritization for immediate resuscitation and care areas with both limited capacity and medical resources.

As stated above, the purpose of this effort was design of a simple system for representing and capturing treatment and resource allocation decisions by triage managers and translating those decision data to a limited number of dimensional attribute scores describing the DMs. This reduction made it possible to consolidate this representation of the data representing a group of DMs in a single multi-dimensional index.

Triage Manager Attribute Development. The first step in the process was derivation of a set of dimensions representing the characteristics along which triage DMs could be expected to vary. Based on consultation with Dr. Irizarry, subject matter expert with experience as an active duty Army Colonel who previously served as NATO Force Surgeon, and who helped draft original TC3 doctrine for the Army, the team arrived at the set of *decision-maker attributes (DMAs)* outlined in Table 1.

These included *Denial*, the willingness to make triage decisions involving the denial or withdrawal of care. It may be assumed that these denial decisions would be based on the DM's understanding of the total situation, the patient's prognosis relative to the resources available, and other factors that would come into play that would cause a triage DM to conclude that the best course of action may be to withdraw or deny treatment to a given patient. The second attribute defined was *Policy*, or the willingness to deviate from policy or doctrine when making treatment and resource allocation decisions. Examples of such decisions and the reasons behind them would include a decision to treat patients in an order that contradicts operational orders or commanders' intent, decision to provide care to enemy combatants, or a decision to provide treatment to patients who might be judged to have very little chance of survival. The third attribute was *Quality*, or quality of life prioritization. Those high on this attribute would be more willing to consider the patient's likely future quality of life when deciding whether to treat or how to prioritize the needs of patients with injuries with different long-term impact on the patient's quality of life, such as those with first degree burns over 90% of their body surface area. A fourth attribute was *Mission*, or prioritization of mission success, in determining the order of treatment for patients. Those high on this attribute would be more likely to treat patients in leadership positions or of greater tactical importance before patients with more serious injuries, or who would otherwise be likely to be treated earlier absent those positional differences. The fifth attribute defined was *Value*, or perceived differences in the value of different lives. DMs high on this attribute would be likely to make treatment order or provision decisions based on information about who the patients are, rather than what their injuries are. Enemy combatants or children would be examples of patients whose lives might have different value to the DM than the warfighters in the DM's unit.

Table 1. Decision Maker Attributes (DMA)

Number	DMAs	Definition	Scale
1	(Denial) Willingness to make decisions involving denial or withdrawal of care	Based on the totality of the decisions made by the DM that indicate the DM's willingness to make denial/withdrawal of care following assessment.	1=low willingness, 10=high willingness
2	(Policy) Willingness to deviate from policy	Based on the totality of the decisions made by DM that indicate the DM's willingness to deviate from policy or doctrine when making treatment and medical resource allocation decisions.	1=unwillingness to deviate, 10=very willing to deviate
3	(Quality) Quality of life prioritization	Based on the totality of the decisions made by the DM to indicate the degree to which the DM considers the survivors' likely future quality of life in decision making, as opposed purely to their odds of survival.	1=low priority assigned to quality of life for survivors, 10=high priority assigned to quality of life
4	(Mission) Prioritization of mission success	Based on the totality of the decisions made by the DM to indicate the degree to which the DM prioritizes mission success over saving as many casualties as possible regardless of the team's ability to continue to persecute the mission.	1=low priority, 10=high priority
5	(Value) Perceived differences in value of different lives	Based on the totality of the decisions made by the DM to indicate the degree to which the DM will change his model of treatment based on characteristics of the casualties, such as whether some are children, or older people, or whether some patients are personally known by the DM.	1=treats all lives as equally valuable, 10=puts a higher priority on some lives than others

Creation of Triage Scenarios. Triage scenarios were designed using *probes* to add complexity and variation to the situations to be perceived by the DMs, intended to represent the kinds of decision-relevant variation likely to be encountered in real world TC3 scenarios. A preliminary list of the probes to be included in scenario design is listed in Table 2. These probes were used as the basis for defining an initial set of scenarios intended to differentiate DM respondents based on the content of the scenarios and details about the casualties. Table 3 provides an example of one of 3 scenarios designed for this purpose. Casualty details, injuries, and treatment options were defined using standard reporting conventions and treatment options listed on the Joint Trauma System (JTS) Point of Injury TC3 After Action Review (AAR) form (JTS 2020; see Figure 1).

Table 2. TC3 Scenario Probes by Category.

1. Patient
   - Injury set
   - Injury severity
   - Injury quantity
   - Injury onset
   - Number of patients
   - Patient type (e.g., pediatric)
   - Injury progression
   - Others contributing factors leading to survivability
   - Necessity for mission
2. Provider
   - Provider skills
   - Number of providers
3. Resources
   - Immediate resource availability
   - Long-term resource availability/sustainment
4. Time constraints
   - Need for quick treatment
5. Emotions
   - Life/death choice
   - High stress
6. Military environment
   - Tactical situation
   - Mission
7. The Unknowns: Other mission relevant factors about which no information is available

For each scenario, respondents were presented with a background vignette, providing an explanation of where the casualty event takes place, who the casualties are, and what resources and equipment the respondent had available. Respondents were informed that each treatment selected for a given patient would take a fixed period of time, which varied by treatment and was visible to the respondent. After reading the initial vignette, respondents were asked to determine which patients to examine in what order, and which treatment(s) to provide to those patients. Each examination accounted for 30 seconds of patient treatment time. Upon selecting "examination" for a given patient, the respondent was presented with a patient assessment upon which treatment decisions could be based (see the last 6 rows of Figure 2 (left) for examples). Following examinations of some or all patients, respondents were asked to decide which patients receive which treatments, chosen from drop-down lists in which order, through a single round of treatment decisions. Time requirements for examination and all treatments were determined by a medical SME. The interface the respondent used to select treatments and track total elapsed time is presented below in Figure 2. Respondents were also asked to self-report their levels of medical training, provided below in Table 4.

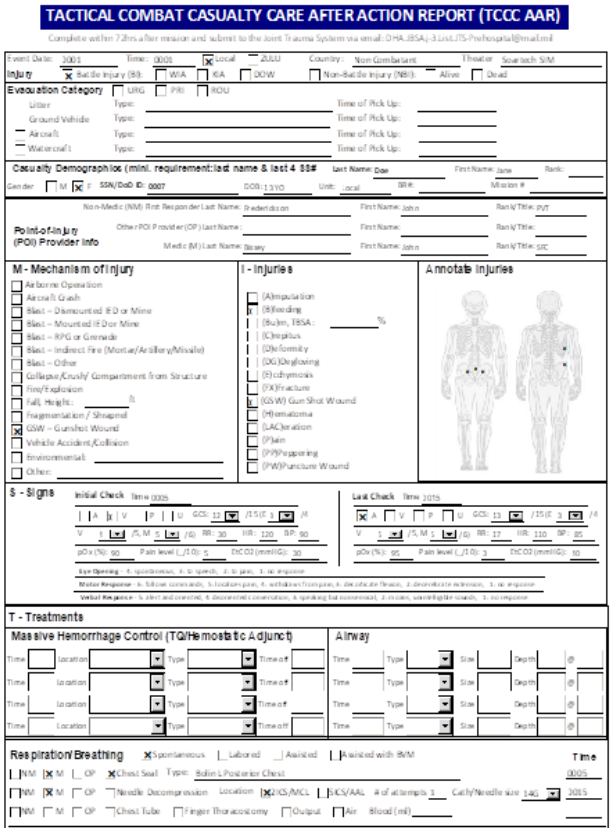


Figure 1. Point of Injury Joint Trauma System (JTS) TC3 AAR form (https://jts.amedd.army.mil/assets/docs/forms/POI_TCCC_AAR.pdf)

Respondents. A total of 10 respondents provided treatment plans through a web-based interface, role-playing as triage managers for this pilot evaluation. Based on self-reported medical training, the respondent group consisted of 2 MDs, 1 Nurse, 1 EMT, I Paramedic, and 5 laypeople without medical training (utilized by necessity due to available time to complete the design evaluation). Each respondent was asked to evaluate 3 scenarios. One respondent (the nurse), was only able to complete 1 of 3 scenarios, yielding a total of 28 DM scenario evaluations. Data reported here are synthetic near approximations of the scenario responses and DMA values assigned to these pilot effort contributors: No actual human subjects data is reported in this document.

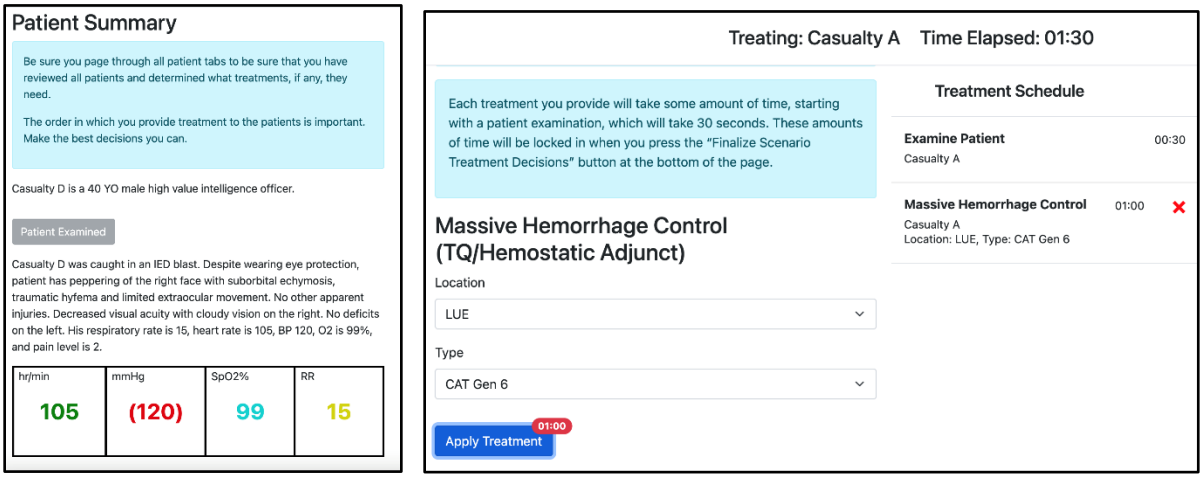


Figure 2. Interface displaying vignette treatment options for Casualty A in Scenario 2 (left), and summary of information presented to respondents following selection of the "Examine Patient" option (right)

Table 3. Example Scenario and Casualty List.

Initial Setting	A squad of 8 soldiers and a high value intelligence officer are in two vehicles transiting from a village to a base in a desert region when there is an explosion presumed to be an IED approximately 90 km from the base. There do not appear to be any hostiles nearby. Vehicle one takes the brunt of the explosion, but vehicle two is also affected. Vehicle two still has a working radio.
Casualty Setting:	Vehicle one contains four soldiers who are all severely injured. One of the soldiers and the high value intelligence officer in vehicle two are also injured by flying glass and shrapnel. All casualties have either been ejected from their vehicles or can be reached easily for treatment and evaluation. Both vehicles are within 20 meters of the provider.
Environment:	Daytime in a dry desert region. Evacuation is expected to be available within 40 minutes. The three uninjured soldiers include the medic and two soldiers with no medical training
Mission requirements:	The mission was to get the high value intelligence officer back to base. This mission is of critical importance and needs to be completed.
Casualty A	Initial Information: Casualty A is a 22 YO Male rifleman in vehicle suspected of being hit by an IED. Upon Examination: Initial injuries appear to be 2nd and 3rd degree burns over all extremities and majority of the torso and face, with peppered dirt and shrapnel on all extremities (BSA covers 90% of his body). 1 cm puncture wound left neck extending into mouth and jaw. Open Left lower arm extremity fracture with deformity. Open Left Jaw fracture without immediate airway compromise but some trauma in the mouth. Pt is unconscious. His Glasgow Coma Scale is 3. His respiratory rate is 22. His heart rate is 110. His blood pressure is unknown. His O2 is 95%. His pain level is unknown.
Casualty B	Initial Information: Casualty B is a 25 YO male rifleman and squad leader in vehicle suspected of being hit by an IED. Upon Examination: Initial injuries show 2nd and 3rd degree burns of the left half of his body (BSA is 50%), with peppered dirt and shrapnel over the same area. Glasgow Coma Scale is 3. His respiratory rate is 18. His heart rate is 100. His blood pressure is 80. His O2 is 98%. His pain level is 6.

Casualty C	Initial Information: Casualty C is a 36 YO male 36 YO Sergeant struck by shrapnel in Left Arm, through the biceps from an IED. Upon Examination: Entry wound present but no exit wound. Diminished pulses in the Left lower extremity. Pain initially was mild but increasing over time. Swelling noted in the LUE at the humerus. Possible fracture of bone. Pain increasing with worsening of function of the left hand and wrist and numbness noted over time. Patient can ambulate. His Glasgow Coma Scale is 13. His respiratory rate is 15, heart rate is 105, BP 120, O2 is 99%, and pain level is 8.
Casualty D	Initial information: Casualty D is a 40 YO male high value intel officer caught in an IED blast. Upon Examination: Despite wearing eye protection, patient has peppering of the right face with suborbital echymosis, traumatic hyfema and limited extraocular movement. No other apparent injuries. Decreased visual acuity with cloudy vision on the right. No deficits on the left. His respiratory rate is 15, heart rate is 105, BP 120, O2 is 99%, and pain level is 2.
Casualty E	Initial Information: Casualty E is a 26 YO male caught in a vehicle explosion. Upon Examination: Immediate partial amputation of RLE. Pain in right hip and pelvis. Patient's mental status and vital signs are deteriorating very rapidly. His respiratory rate is 25, heart rate is 110, BP 90, O2 is 95%, and pain level is 10.
Casualty F	Initial Information: Casualty F is a 22 YO male caught in a vehicle explosion. Upon Examination: He has sustained a shrapnel wound to his left chest and is having difficulty breathing. His respiratory rate is 18, heart rate is 110, BP is 120, O2 is 99%, and pain level is 3.

The number of casualties included in each scenario was as follows:

- Scenario 1: 2 casualties, with amputations and chest wounds
- Scenario 2: 6 casualties, varying ranks and value to Scenario
- Scenario 3: 5 casualties, including bystanders and an enemy combatant.

As described above, respondents were instructed to assemble a treatment schedule through a single round of observation and treatment. Figure 3 provides a partial list of the treatment options made available to respondents based on the standard JTS POI TCCC AAR form (JTS, 2020).

Feature Specification. Treatment choices were modeled as a series of vectors corresponding to the choices that respondents made across the 28 scenarios evaluated. This initial dataset included a total of 40 vectors qualitatively evaluated for patterns of interest and relationships to DMA scores.

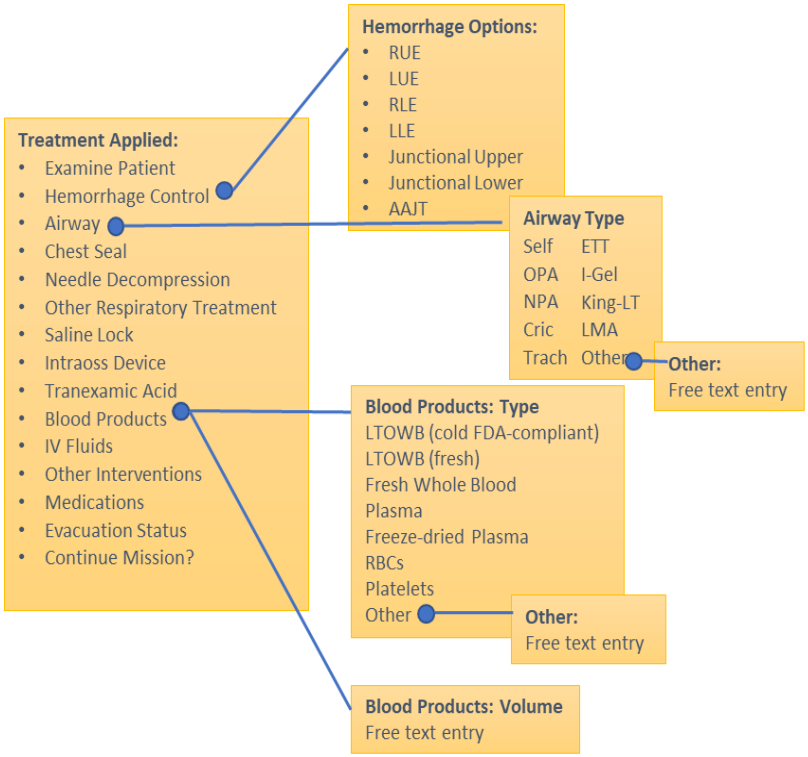


Figure 3. Partial List of Treatment Options Available Through the DM Treatment Interface (derived from the JTS POI TCCC AAR)

DMA Score Assignment. Following completion of treatment plans across 28 scenarios by the 10 respondents, one of the authors reviewed the raw data and assigned scores on the 5 DMAs listed in Table 1 to the evaluator for each scenario evaluated. Thus, 9 of the 10 respondents received 3 different sets of DMA scores, one for each scenario they evaluated. These are reported in Table 4, along with mean DMA scores by respondent, means and SDs for the total sample of 28 respondent scores assigned and average DMA scores for those respondents with medical training versus those without. Statistical comparisons were not performed on these data.

Table 4. DMA Respondent Means and Overall Descriptives (k = 10, N = 28)

Respondent	Level of Training	Denial	Policy	Quality	Mission	Value
1	Layperson	3.67	1.67	2.67	7.33	1.33
2	Layperson	3.33	2.00	3.67	7.00	1.67
3	Layperson	1.33	3.67	2.00	3.33	1.67
4	Physician	6.00	1.33	6.00	5.67	4.67
5	Physician	5.67	2.00	5.33	6.67	4.67
6	Paramedic	5.00	2.00	5.33	6.67	4.00
7	Layperson	4.67	4.67	4.33	8.00	6.33
8	EMT	7.50	3.00	3.00	5.67	5.00
9	Layperson	4.00	3.00	3.33	8.33	3.33
10	Nurse Practitioner	4.00	3.00	4.00	5.00	3.00
	Mean	4.42	2.58	4.04	6.46	3.50
	SD	2.48	1.53	1.93	1.79	2.32
	Layperson Mean	3.40	3.00	3.20	6.80	2.87
	Medical Mean	5.63	2.27	4.73	5.93	4.27

Note: Respondent 10, Nurse Practitioner, only evaluated one scenario. All other respondent DMA means are based on separate ratings across 3 scenarios for each respondent.

Visualization of DMA Scores

The next challenge was an evaluation of the practicality of consolidating the information represented across all 5 DMA variables into a single visualization. The data described above in Table 4 were evaluated across all 15 possible pairings of these 5 variables using Gaussian Mixture Modeling to detect clusters. The cluster size and densities across all these pairings were used to create a set of alignment height maps. Figure 4, below, is an illustration of 5 of these 15 pairings. In each graph below, the vertical axis indicates the relative frequency with which DMA scores clustered by pairings of dimension. This tells us how our DMs' DMA scores tended to cluster across all 28 scenario evaluations performed. While viewing these data as sets of pairs of DMA variables keeps the data readable by human eyes, further consolidation was necessary, as described next.

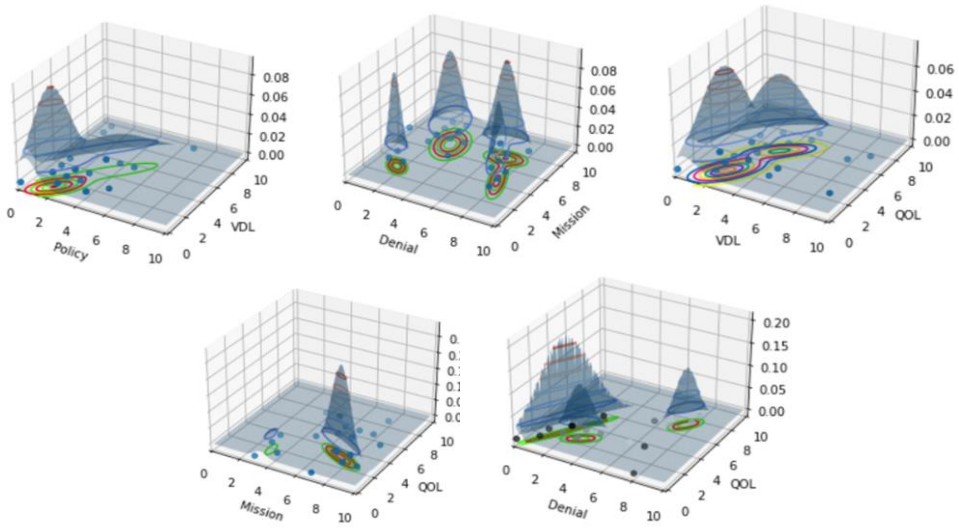


Figure 4. Five of 15 Gaussian Mixture Model solutions for DMA pairings

Note: QOL: Quality of Life; VDL: Value of Different Lives (see DMAs definitions in Table 1)

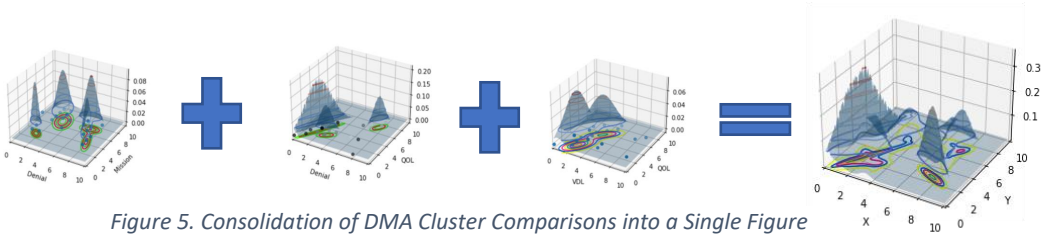


Figure 5. Consolidation of DMA Cluster Comparisons into a Single Figure

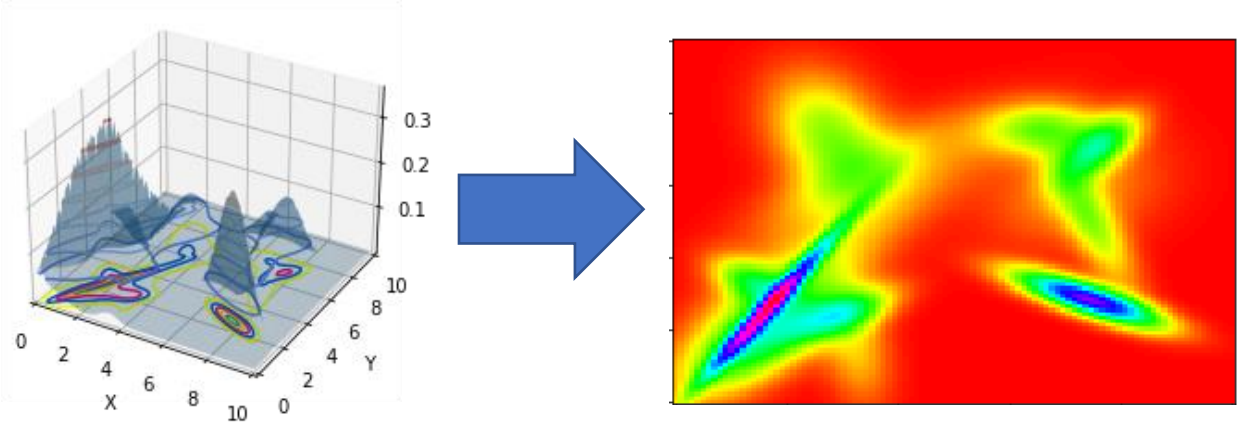


Figure 6. Depiction of DMA Scores Consolidated into 3D Composite View (left) and 2D Heat Map (right)

It was more useful to consolidate these DMA pair graphs into a single graphical representation of how DMAs cluster across the entire set in a single visualization. This is depicted in Figure 5, and provides a representation of how any set of human DMs' standing on any set of DMAs can be depicted. The data are still more useful, however, if converted from the composite 3D terrain view into a two-dimensional depiction that can provide a view of DM clustering and alignment on DMAs at a glance, as depicted in Figure 6. Figure 6 depicts this same information as a heatmap, where the color red indicates zero density of data. The data in this heatmap is a two-dimensional representation of our 10 participants' DMA scores across the 28 evaluated scenarios. This representation can be used to identify two different types of misalignment, both of which are illustrated by the "aligned" and "misaligned" dots in Figure 7. Misalignment, as so indicated, can represent both:

- *Human DM outliers*: as indicated by the degree of convergence or separation of any single DM from his or peers on the DMA dimensions graphed.
- *Misaligned algorithm DMs*: The degree to which the DMA convergence plots for an algorithmic DM fail to overlap with the clustered (non-red) areas will indicate a lack of alignment between the DMA values of the algorithmic DM and the human reference DM population used to generate the heat map.

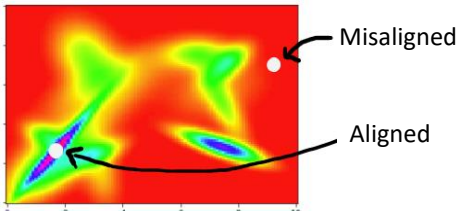
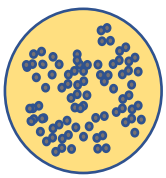


Figure 7. Illustration of How a DMA heatmap can be used to quantify the alignment of human or algorithmic DMs with a reference DM population



Dots represent clusters across all pairings of DMA dimensions. Density of the dots indicates convergence among human DMs on DMA dimensions.

Figure 6. Conceptual Representation of DMA Clustering for a Population of Human DMs

The separation between these points can collectively be measured as Euclidean distances in n-dimensional space (Gordon, 1999), as discussed by Halibisky (2022). While the heatmap depicted in Figures 6 and 7 is a depiction of the actual data collected as part of this IR&D effort, we can use the same approach to conceptualize the set of any consolidated DMA values for a set of human reference DMs across scenarios using a similar image, as depicted in Figure 8.

Such a depiction of human DM standing on a combined set of DMAs can be used as the basis for comparison to the DMAs of an algorithmic DM, as depicted in Figure 9. Figure 9 depicts the conceptual space representing a reference set of human DMs' standing on a set of DMAs in yellow on the right. A single hypothetical algorithmic DM, evaluating the same or a similar set of scenarios, would generate corresponding values on all the same DMAs, as represented by the set of DMAs in the blue circle on the left. The degree of overlap between the DMA values for the algorithm and the reference sample of human DMs could be quantified as the degree of convergence or alignment between the algorithm and that human reference DM group. This conceptual approach will be extremely useful in the future evaluation of how closely triage decisions made by algorithms match those made by human reference DMs.

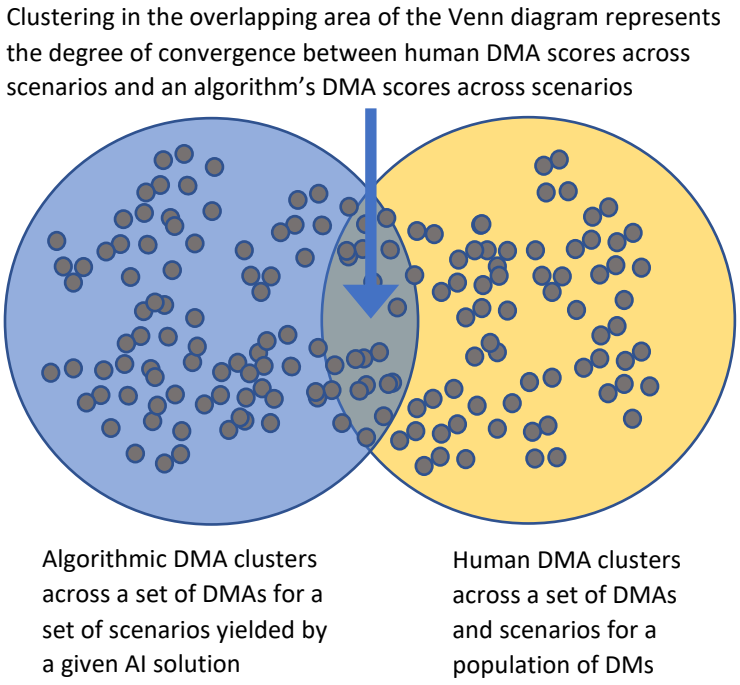


Figure 7. Conceptual Depiction of the Overlap between DMAs Yielded by a Human Population for a Set of Triage Scenarios with the DMAs Yielded by a Hypothetical AI Algorithm

Additional Exploratory Analyses

The next challenge associated was an evaluation of the degree to which the DMA variables so assigned could be determined to have content-valid associations with the treatment decisions. The first challenge was the exploration of the relationship between the highly varied structure of the DM treatment choices, including the relationships between provider training level, which patients were treated, order in which that treatment was provided within scenarios, time spent per patient, and which specific treatments were provided, including those containing open-ended response data. For this, the team used a commercial tool called HiPlot.

HiPlot Analyses. This tool is designed for visualizing hyperparameters in machine learning, and can be of use in visualizing the complexity of relationships among uniquely structured predictors and outcome variables of interest. Figure 3 below provides a visual depiction of the entire set of treatment features data in relation to *Quality* DMA scores for the 28 scenario evaluations. In these "spaghetti" plots, each of the 28 "strands" represents one RDM scenario run. While the figure below is too dense to be illuminating, more narrowly tailored sets of these data yielded interesting findings such as:

- DMs with similar *quality* scores made similar choices for patient treatment order
- DMs with a low *denial* score appeared to spend more time per patient than DMs with a high *denial* score
- DMs with medical training had higher *denial* scores than DMs with no medical training.

It must be observed that given the limited nature of these data, it is not the findings themselves that are important (as these are highly unlikely to generalize beyond this limited dataset), but the confirmation that the HiPlot tool is useful for detecting these patterns that could be missed using other exploratory analytics.

C**ausalNex Analyses**. CausalNex is a hybrid open-source toolkit that uses Bayesian Networks to combine machine learning and domain expertise to model causal reasoning. It relies on the NOTEARS Algorithm (Zheng, Aragam, Ravikumar, & Xing, 2018) as a method to learn structures and understand conditional dependencies between variables. This allows domain knowledge to refine model relationships found by machine learning, and builds predictive models

based on structural relationships. CausalNex was used to identify the most important relationships between DMA variables and scenario probes/features. While not explicitly hypothesized, expectations were that DMA variables would exhibit causal relationships with DM features including

- order of patient treatment
- time spent per patient
- total time spent by the DM
- specific treatments provided

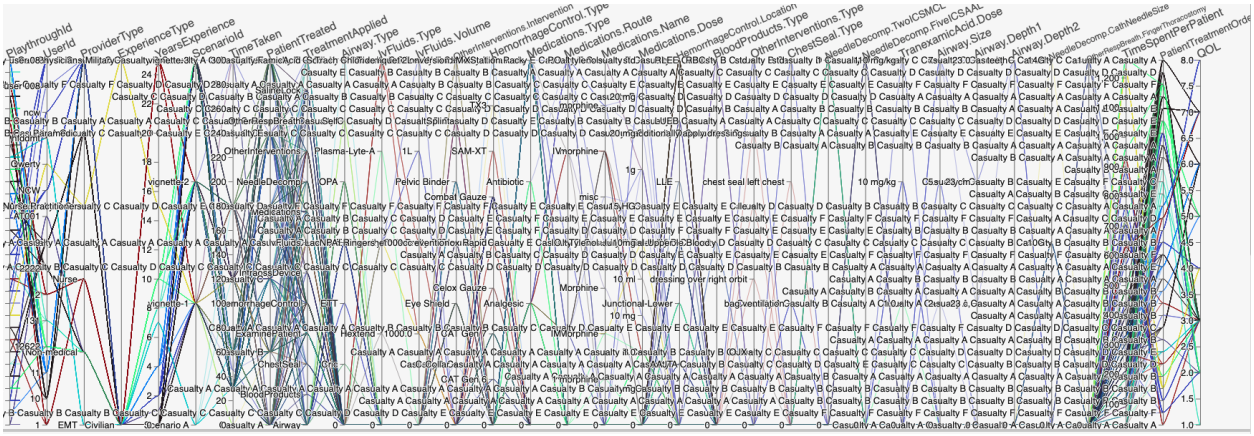


Figure 8. Set of treatment decision features data in relation to all Quality DMA scores assigned

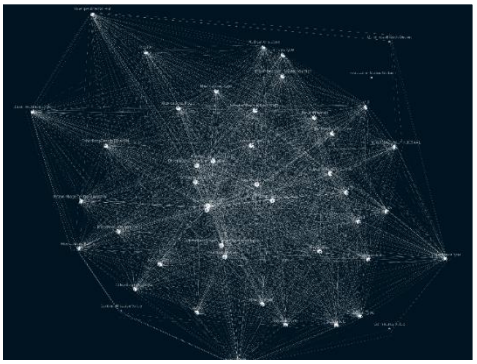


Figure 9 Original Model for N=28 Scenarios Prior to Edge Reduction

Figure 11 depicts the original model yielded by CausalNex analyses prior to any edge reduction introduced based on strength and plausibility of relationships. By contrast, Figure 12 depicts an updated model following an initial round of edge reduction. This significantly reduced the number of edges and nodes in the model, and made it possible to visually detect a pattern of findings consistent with expectations regarding the relationships of probes/features with DMA variables.

Note that the probes/features at the center of the network include total time taken, years of experience, patient treatment order, specific treatment applied, and patient treatment order. These are all surrounded at similar intervals by the 5 DMA variables included in this project: *mission*, *denial*, *policy*, *quality*, and *value* (as listed in Table 1).

Why This Matters. This is an illustration that even using a set of overly simplistic triage scenarios and an extremely small dataset, it was possible to use CausalNex to identify possible causal relationships involving triage decision nodes and/or DMA variables.

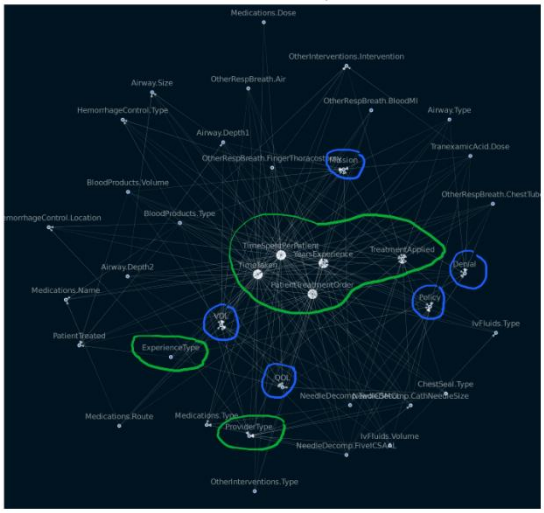


Figure 10. Model Following Edge Reduction for N = 28 Scenarios. Relevant probes/features are circled in green, while DMAs are circled in blue.

Discussion

This paper described a feasibility assessment of the degree to which triage scenario management decision-making could be captured and translated to a set of attributes describing the decision-maker. Goals included the translation of these attributes to a two-dimensional display, facilitating comparisons between aggregated human decision-making and AI systems to be developed for complex decision-making. The effort also explored the feasibility of using data visualization and causality evaluation tools to explore relationships between specific treatment decision-points and DMAs. The effort yielded the following accomplishments.

- Generated a preliminary scheme for medical triage scenario generation
- Demonstrated an approach for consolidating DMA values into a single heatmap
- Demonstrated how this heatmap can be used to assess the alignment of algorithmic DMs to a group of human reference DMs
- Illustrated the practicality of two data causality and visualization tools, HiPlot and CausalNex, for the examination of predictive relationships of triage treatment variables with DMAs.

It must be noted, however, that this effort was of extremely limited scope. Expansions and replication of this work must take into account the following limitations and development considerations.

DMA Score Assignment and Analyses. Since the primary purpose of this effort was the evaluation of a scenario design tool and data visualization approach, steps such as multiple rater evaluation and computation of within-respondent variance in DMA scores were not evaluated under the current effort. In order to generate usable data, a replication of this effort would need to incorporate a far more rigorous DMA score assignment process, including multiple raters, and analyses that included correction for within-respondent variation across scenarios.

Theoretical foundations for DMA dimension development. Any replication of this effort should begin with development and evaluation of DMA dimensions based on a sound decision-making theory, as well as a cluster- or factor analytic evaluation of both the hypothesized attributes as well as any that emerge from the respondents' treatment plan data.

Need for multiple-round triage scenario design. The scenarios developed for this work were based on a *single round* of caregiver decisions, with fixed time costs assigned to each treatment option. Real world triage decision-making is far more challenging specifically because it requires so many successive decisions based on best available information. Triage decision-making involves multiple rounds of patient evaluation, changes to patient status, and finite resources that are exhausted. The extended nature of the scenario is the most important part of what makes it challenging. Realistic, multiple-round modeling of POI TC3 treatment decision-making, even for events with a small number of casualties, is a critical necessity. Extension of this approach to the capture of data representing more complex patient scenarios – such as those at a surgical care center where number of caregivers and patients would be far higher than in a POI scenario – would make multiple-round scenario modeling even more important.

Need for cluster analyses to translate raw decision data into DMA scores. One implication of this is the exponential increase in complexity of the decision data. The number of alternative paths available to triage decision-makers reviewing the status of the same patients multiple times within a scenario, with different survival outcomes for different DMs across rounds of treatment, will involve thousands of alternative paths – and will result in data with vastly different structures across DMs. *As the complexity of the decision data from DMs increases, the importance of a cluster-analytic based approach to translate the raw decision data to a manageable number of theoretically based DMAs increases dramatically*. This IR&D did not explore this problem, as it was beyond the scope of the current effort, but future work translating DM treatment decisions to DMAs must take this into account.

Impact of Vignette-Based Scenarios. The artificiality of modeling triage decision-making using a vignette-based policy capturing approach (Cooksey, 1996; Johnson & Raab, 2003; Sherer, Schwab, & Heneman, 1987) is also likely to induce meaningful differences in the nature of the decisions made from those that would be made under real-world conditions where those decisions include specific treatments provided by the DM in real time rather than options selected from pull-down menus or typed into fields on a self-report form.

Event-based scenario design. The scenarios used in this work were developed based on the JTS POI TCCC AAR form (JTS, 2020), which provides conventions and standard expectations for the evaluation and treatment of patients at POI. They did not, however, incorporate many additional decision-relevant considerations that would affect triage decision-making in operational settings, such as the tactical environment, complex timelines for supply and resupply of medical resources, changes in availability or status of treatment staff, among other considerations. Any replication or expansion of this effort should be designed using an event-based model (Fowlkes, Dwyer, Oser, & Salas, 1998; Oser, Gualtieri, Cannon-Bowers, & Salas, 1999) to generate scenarios that will translate better to the nature of real-world triage decision-making.

REFERENCES

Achkoski, J., Koceski, S., Bogatinov, D., Temelkovski, B., Stevanovski, G., & Kocev, I. (2017). Remote triage support algorithm based on fuzzy logic. BMJ Military Health, 163(3), 164-170.

Andronie, M., Lăzăroiu, G., Iatagan, M., Uță, C., Ștefănescu, R., & Cocoșatu, M. (2021). Artificial intelligence-based decision-making algorithms, internet of things sensing networks, and deep learning-assisted smart process management in cyber-physical production systems. *Electronics*, *10*(20), 2497.Cooksey, R. W. (1996). Judgment analysis: theory, methods, and applications. San Diego: Academic Press.

Fowlkes, J., Dwyer, D.J., Oser, R.L., & Salas, E. (1998) Event-based approach to training (EBAT), The International Journal of Aviation Psychology, 8:3, 209-221, DOI: 10.1207/s15327108ijap0803_3

Gordon, A. D. (1999) Classification, 2nd Ed. Chapman & Hall: Boca Raton.

Halibisky, B. (2022). Euclidean distance in n-dimensional space. Online at: https://hlab.stanford.edu/brian/euclidean_distance_in.html

Hopko, S. K., Mehta, R. K., & McDonald, A. D. (2021). Trust in automation: Comparison of automobile, robot, medical, and cyber aid technologies. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting (Vol. 65, No. 1, pp. 462-466). SAGE Publications.

Johnson, J. G. & Raab, M. (2003). "Take The First: Option-generation and resulting choices" Organizational Behavior and Human Decision Processes. 91 (2): 215–229. doi:10.1016/S0749-5978(03)00027-X.

Joint Trauma System Clinical Practice Guideline 11 (18 Sep 2020). Documentation Requirements for Combat Casualty Care. https://jts.amedd.army.mil/assets/docs/cpgs/Documentation_Requirements_for_Combat_Casualty_Care_18_Sep_2020_ID11.pdf

Klein, G. A., Orasanu, J. M., Calderwood, R., & Zsambok, C. (1993). Decision Making in Action: Models and Methods. Ablex. ISBN 978-0-89391-943-6.

Klein, G. A. (2008). "Naturalistic Decision Making". Human Factors: The Journal of the Human Factors and Ergonomics Society. 50 (3): 456–460. doi:10.1518/001872008X288385.

Oser, R.L. Gualtieri, J.W. Cannon-Bowers, J.A.. & Salas, E. (1999). Training team problem solving skills: an event-based approach. Computers in Human Behavior, 15: 3-4, 441-462, ISSN 0747-5632, DOI 10.1016/S0747-5632(99)00031-X.

Selten, R. (1999) What is Bounded Rationality? SFB Discussion Paper B-454, Proceedings of the Dahlem Conference 1999.

Sherer, P. D., Schwab, D. P., & Heneman, H. G., III. (1987). Managerial salary-raise decisions: A policy-capturing approach. Personnel Psychology, 40, 27-38.

Simon, H.A., 1957, Models of Man, New York, Wiley & Sons.

Stevanoski, G., Kocev, I., Ackoski, J., Koceski, S., & Temelkovski, B. (2016). Implementation of a system for physiological status monitoring by using tactical military networks. Defence Science Journal, 66(5), 517-521.

Todd, P. M. & Gigerenzer, G. (2001)."Putting naturalistic decision making into the adaptive toolbox" Journal of Behavioral Decision Making. 14 (5): 381–383. doi:10.1002/bdm.396.

Zsambok, C E. & Klein, G. A. (1997). Naturalistic Decision Making. L. Erlbaum Associates. ISBN 978-0-8058-1874-1.

Zheng, X., Aragam, B., Ravikumar, P., & Xing, E.P. (2018). DAGs with NOTEARS: Continuous optimization for structured learning. Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS), Montreal, Canada. https://proceedings.neurips.cc/paper/2018/file/e347c51419ffb23ca3fd5050202f9c3d-Paper.pdf