**Position: We Can Measure XAI Explanations Better with Templates**

[**http://ceur-ws.org/Vol-2582/paper2.pdf**](http://ceur-ws.org/Vol-2582/paper2.pdf)

<0. Basics>

|  |  |
| --- | --- |
| ExpG | Hoffman의 **Explanation Goodness** |
| ExpS | Hoffman의 **Explanation Satisfaction** |

**[Contextualization, Actor, and Timing]**

|  |  |  |
| --- | --- | --- |
|  | ExpG (Explanation Goodness) | ExpS (Explanation Satisfaction) |
| Contextualization | Task와 관련되어 정의되지 **않는다**. | **Task와 관련**되어 정의된다. |
| Actor | **연구자의 관점**에 초점을 둔다. | 사용자가 **task를 수행하는 것**에 초점을 둔다. |
| Timing | **언제든지 측정**될 수 있다. | Task에 관련되어 정의되므로 **task가 종료된 후에 측정**된다. |

**[Background, Tasks, Benefits, and Scope]**

|  |  |
| --- | --- |
| Background | 현재의 많은 연구들은 **ExpS에 집중**하고 있다. |
| Tasks | **ExpS**가 **더 쉽게 조작할 수 있지만 많은 experimental noise**를 유발한다. |
| Benefits | **ExpS**의 유용성은 사용자의 **제한된 시스템 노출에 의해 방해**된다. |
| Scope | **ExpG는 더 큰 범위의 행동**을 할 수 있게 하고, **ExpG** 메커니즘이 **explanation template에 대한 이유 설명**에 적합하게 한다. |

<1. Background: Most Current Research Has Focused on ExpS>

|  |  |
| --- | --- |
| ExpS | 설명의 효과를 조절할 때 설명의 **soundness와 completeness**를 이용한다.  **Soundness:** 진실된 것 외에는 없다.  **Completeness:** 모든 진실을 설명한다.   * 이것들은 binary property가 아니라 **smooth continuum**이다. 따라서 완전한 soundness 또는 completeness는 불가능하다. |
| ExpG | **설명이 얼마나 자세한지**를 고려한다.   * “The explanation of the [software, algorithm, tool] is sufficiently **detailed**.” |

<2. Tasks: ExpS is “Easy” to Operationalize, but Noisy>

4개의 experiment treatment가 있는 실험

|  |
| --- |
| 1. **어떤 treatment도 clear winner가 아니다**.  2. 어떤 결정은 **설명이 없음에도 불구하고 모든 참여자가 잘 한** 경우가 있다.  3. **설명이 있음에도 불구하고 AI의 다음 행동을 예측하지 못한** 경우가 많다.  4. **learning effect에 대한 증거가 없다.** (decision point가 시간이 지남에 따라 sequential함) |

XAI 연구에 대해 다음과 같은 도전 과제가 있다.

|  |  |
| --- | --- |
| 실험 참여자의 **과제 수행 능력** | 과제 수행 능력(AI의 다음 행동을 예측)이 **수많은 다른 것들의 영향을 받는다.**   * **인식의 필요성, 과제의 흥미도, 도메인 경험** 등 * 과학에 대한 질문을 다른 과학자에게 할 때와 같은 사람이 같은 질문을 어린이에게 할 때를 비교 |
| **State/action space**에서 **variability**가 큼 | **어떤 선택은 쉽지만 반대로 어떤 선택은 매우 어려워진다**.   * **쉬운 선택:** **100%에 가까운 예측 정확도**를 보이는 것을 발생시키는 것 * **어려운 선택:** **0%에 가까운 예측 정확도**가 나타나는 것 * 따라서 이러한 **floor and ceiling effect에 의해 treatment 간의 어떤 variation은 모호**해진다. |
| 실험 참여자의 **예측에 대한 partial credit** 할당 | AI는 Figure 1에서보다 **더 큰 action space**에서 사용되므로 실험 참여자가 올바른 것을 선택할 확률은 극도로 낮아지고, 따라서 **어떤 답이 다른 답보다 더 고려될지를 생각하는 것**이 자연스럽게 보인다.   * 이를 위해서 실험 참여자는 **action space 또는 value space의 similarity**를 고려해야 하지만 **양쪽 모두 엄밀히 설계하기 어렵다**. |

<3. Benefits: ExpS’s Usefulness is Hampered by Limited Exposure>

|  |
| --- |
| In-lab User Study가 2-hour window에서 진행되었다고 할 때, 실험 참여자들의 시스템 노출은 꽤 낮다. **Figure 1에서는 limited exposure 때문에 learning effect가 발생하지 않았다**고 할 수 있다.  ExpS에 의한 평가를 할 때 실험 참여자에게 시스템을 노출시키는 것에 대한 또 다른 도전 과제가 있다. (**어떤 decision point를 실험 참여자에게 보여줄 것인가?**)   * 테스트 케이스를 선택하는 한 가지 방법은 **어떤 decision point의 문제를 해결할 것인지**에 대한 접근이고, 이것은 각 상태의 criticality를 측정하고 **agent가 가장 영향이 클 것으로 인식되는 것들을 선택**하는 것이다. |

<4. Scope: ExpG Can Consider a Wider Range of Behaviors, via Templates>

|  |
| --- |
| **Explanation template:**  여기서는 다음 그림과 같은 **Case-based Explanation**을 사용한다.    **왼쪽 그림:** explanation template을 나타내며, 검은색 텍스트는 고정되어 있고 **빨간색 텍스트는 입력에 대한 계산 결과를 바탕**으로 나타난 것이다.   * **ExpS의 결과가 입력에 따라 어떻게 다양하게 나타나는지**를 설명한다.   **오른쪽 그림:** matching percentage의 히스토그램으로, **2개의 classifier가 어떻게 서로 다르게 행동했는지**를 나타낸다.   * Case-based explanation이 **self-refute하는 경우 (낮은 %)**, 또는 **어떤 claim도 없는 경우 (binary 기준으로 50%에 가까움)**와 같은 결과들이 나타났다. * 이것은 ExpS formulation의 사용자들은 **적은 수의 설명만 이용하고 self-refuting은 드물기** 때문에 관측할 수 없다. |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Explanation template을 고려하는 이유:**  ExpG 메커니즘이 어떤 explanation template과 목적에도 사용할 수 있지만, **case-based explanation template**을 사용하는 것도 고려할 수 있다.   * 이런 explanation을 방법 중 하나는 **입력의 가까이에 있는 training example**을 찾는 것이다. * **Explanation Generator를 모든 test set에 대해서 동작**시키면 matching percentage에 대한 histogram을 만들 수 있다. * **하지만 이것은 explanation soundness 관련 문제를 야기할 수 있다.**  |  |  | | --- | --- | | 많은 instance가 100% 근처에 있음 | 사용자들이 **설득력 있다**고 판단할 만하다. | | 50% 정도의 학습 데이터가 입력 label과 일치 | 특정한 claim을 **입증하지 않는다.**   * **Classifier가 binary**일 때 | | 0% 근처의 instance  **(lack of soundness 발생)** | Self-reputing, 즉 **자신을 논박(refute)**하는 설명이 있다.   * “This instance was **labelled an A** because **all the nearby training examples were B’s**.” | |
| **NEXT PAGE** |

|  |
| --- |
| **PREVIOUS PAGE** |
| 이러한 설명들을 **ExpS**로 평가할 때 그 결과는 **제공된 explanation이 convincing인지 self-refuting인지**에 강하게 영향을 받는다. 그러나 **explanation template에서는 이들 각 경우에서 서로 같고, 입력값만 바뀐다.**   * 따라서 **ExpG는 더 넓은 범위를 고려**할 수 있게 한다고 할 수 있다. * ExpG에서 template은 **가끔 자신을 refute하는 설명을 생성**하고, 이것이 **acceptable한지도 결정**할 수 있다. * **자신을 refute하는 설명을 생성하는 것을 회피하는 대안적인 템플릿**을 생성한다면 **static text와 variable part의 계산법**을 오른쪽 그림처럼 조정할 수 있다.   + 새로운 제안에서는 **match되는 것의 개수만 고려**한다. (입력 label과 불일치하는 nearby training example을 효과적으로 무시 가능)   + 결정적으로 **“자신을 refute하지 않는다.”** 그러나 explanation이 **whole truth를 말하지 않으므로 completeness가 감소**한다. |

<5. Scalability: Many design decisions are only validated with ExpG>

|  |
| --- |
| XAI 설계자는 많은 design decision을 마주한다. 예를 들어 특정 XAI는 원래 **single nearest neighbor**를 이용했는데, 이 접근법에 대해 k의 개수를 지정하여 **k nearest neighbor로 확장**하는 등의 방법을 사용할 수 있다.  **ExpS** 메커니즘으로 평가 가능한 **design decision이 매우 많아서 평가가 실질적으로 불가능**하므로 **ExpG를 이용하여 평가**해야 한다.  “몇 개의 explanation을 연구하는 것이 ExpS-oriented 방법에는 적합하지만 template level에서는 ExpG 방법이 필요하다.”   * 이것은 **엄격한 과학에 의한 XAI 시스템**을 설계할 때 **엄밀하고 평가 가능한 ExpG 메커니즘을 향상시켜야** 한다는 것을 의미한다. |