**A Taxonomy for Human Subject Evaluation of Black-Box Explanations in XAI**

<http://ceur-ws.org/Vol-2582/paper9.pdf>

[00. Taxonomy (분류법) Development]

|  |
| --- |
| **<Taxonomy를 구축하는 2가지 접근법>**  **Conceptual-to-empirical:** 연구자는 **특정 이론이나 모델**에 의한 classification을 제안한다.  **Empricial-to-conceptual:** **경험적인(empirical) 각 경우**에 근거하여 분류법(taxonomy)을 유도하는 접근 방법이다.   * 여기서는 공유된 meta-characteristic과 정의된 ending condition 아래에 있는, **이 두 가지 방법이 결합된, Nickerson이 제안한 iterative한 과정**으로 taxonomy를 개발한다.     1. **Conceptual-to-empirical 접근법**으로 시작한다.   * 이것을 위해서 특정 이론이나 모델에 근거한 **classification이 제시**되어야 한다. * **XAI evaluation을 위한 제안된 카테고리를 통합**하여 classification을 제시한다.   2. 이 방법으로 만들어진 taxonomy는 **ideal(pure) type**를 설명한다. |

[01. Structured Literature Review]

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Empirical-to-conceptual iteration의 한 부분으로, **Structured Literature Review(SLR)을 이용하여 validate하고 taxonomy를 iterate**하려고 한다.   * 이것을 통해 **taxonomy를 이용하여 영역(field)이 어떻게 구성되어 있는지 설명**할 수 있다. * SLR은 Kitchenham과 Charters에 의해 제안된 접근 방법을 따른다.   **<제안된 search strategy>**   |  |  | | --- | --- | | Source Selection | Google Scholar에서의 XAI에 대한 exploratory search에서는 **관련 연구가 여러 출판자들로 분산**되어 있다. | | Search Query | Exploratory search를 통해 **search query를 구성하도록 돕는 관련된 키워드, 동의어, 개념** 등에 대한 기초적인 이해를 할 수 있다.   * XAI와 human subject evaluation 접근법에 대해 **서로 다른 용어**를 사용하고 있다. * 이전 연구에서는 XAI 또는 explainable artificial intelligence라는 용어를 사용하지 않았으므로, **우리의 검색 쿼리는 group과 term들로 구성**되어 있다. * 용어들은 group domain에서 **서로 비슷한 semantic meaning**을 가지거나 **상호 간에 바꿔서** 사용될 수도 있다. * 여기서는 **서로 다른 용어로 표현되는 3개의 그룹의 교집합 (Table 1 참조)**에 대하여 알아본다. | | Study Selection Criteria | 검색 결과를 **6개의 exclusion criteria (EC)와 1개의 inclusion criterion (IC)**로 필터링한다.   * IC-1은 **XAI context에 대한 setup과 human subject의 결과**에 대한 연구이다. * Black-box explanation problem을 해결하는 publication에 대한 조사를 제한한다. | | Study Analysis | 2019년 9월에 **Scopus에서 검색을 수행**하였다.   * 총 653개의 잠재적으로 관련된 publication이 발견되었다. * 현재(2020년 3월) inclusion criterion에 포함되는 publication을 분석하는 중이다. | |

[02. Taxonomy of Human Subject Evaluation in XAI – Task Dimensions]

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Mohseni와 Ragan은 설명에 대한 평가에서 **2가지 종류의 human involvement를 구분**하였다.   * **Feedback setting:** 참여자는 실제 설명에 대한 **피드백을 제공하고 실험자는 이 피드백을 통해 설명의 품질을 측정**한다. * **Feed-forward setting:** 설명이 제공되지 않으며, **사람이 reasonable한 설명의 예시를 생성하고 이것이 설명 알고리즘의 학습 데이터**가 된다.   Doshi-Velez와 Kim은 **task abstraction의 수준에 따라 다른 2가지 종류의 human subject evaluation**을 구분하였다.   * **Application-grounded evaluation:** 실제 **응용 context에서 실험**을 수행하고, 해당 **응용 context에서 설명의 품질을 평가**한다. (테스트 성능을 주로 이용) * **Human-grounded evaluation:** **단순화 또는 추상화된 실험**을 진행하며, target application의 본질을 유지하는 것을 목적으로 한다.   **<Multiple types of user tasks>** - **설명의 품질 정보를 추출**하기 위해 제안됨   * 참여자에게 제공되는 정보와 return에서 요청되는 정보를 이용하여 구분한다.  |  |  | | --- | --- | | Verification task | 참여자는 **입력값, 설명, 출력값**을 제공받고 설명에 대한 **만족도**를 요청받는다. | | Forced choice task | Verification task의 확장 (**여러 개의 경쟁하는 설명** 중에서 선택) | | Forward simulation task | 참여자는 **입력값과 설명**을 제공받고 **시스템의 출력**을 예측할 것을 요청받는다. | | Counterfactual simulation task | 참여자는 **입력, 설명, 출력값**을 제공받고 **대안 출력값(counterfactual)**을 제공받는다. 이때 **입력의 어떤 변화가 counterfactual을 얻기 위해 필요**한지 예측한다. | | “Clever Hans” detection task | 참여자는 **flawed model을 identify 및 debug**해야 한다.  (Naïve 또는 short-signed predictor 등) | | System usage task | 참여자가 **시스템을 사용하고 그것의 대표 목적에 대한 설명을 제공**받는다. (decision-making system 등)   * 설명의 품질은 **decision quality**라는 용어로 평가된다. | | Annotation task | 참여자는 **모델의 입출력에 대해 적절히 설명**한다. |   **<Intended explanation goals>** - **효과적인 evaluation**을 위하여 제안됨   |  |  | | --- | --- | | Transparency | 시스템이 어떻게 동작하는지 설명 | | Scrutability | 시스템의 잘못된 동작에 대해 사용자가 잘못되었다고 알릴 수 있음 | | Trust | 사용자의 시스템에 대한 confidence를 증가 | | persuasiveness | 사용자에게 특정 행동을 하라고 설득 | | 기타 | Satisfaction, effectiveness, efficiency, education, debugging | |

[02. Taxonomy of Human Subject Evaluation in XAI – Task Dimensions]

|  |
| --- |
| Hoffman은 실험 참여자의 XAI 시스템에 대한 이해를 평가하기 위한 **multiple levels of task evaluation**을 설명하였다.   * **Test of satisfaction:** 설명 및 시스템에 대한 인식을 포함한, **참여자의 self-reported 만족도**를 통해 평가한다.   + 여기서 연구자는 참여자가 시스템에 대한 이해를 요구 수준만큼 한다는 것에 대한 확신을 거의 하지 못한다. * **Test of comprehension:** **참여자의 시스템에 대한 mental model 및 이해**를 평가한다. * **Test of performance:** **human-XAI system의 성능**을 측정한다. |

[03. Taxonomy of Human Subject Evaluation in XAI – Participant Dimensions]

|  |
| --- |
| Participant type은 **AI novice (보통 end user), 데이터 전문가(도메인 전문가 포함), AI 전문가**로 구성된다.   * 실험 참여자의 전문성이 **다른 participant-related dimension에 강하게 영향**을 미치기 때문에 이런 distinction은 중요하다. * 실험 참여자의 전문성은 **모집 방법과 참여자의 수**에 의해 달라진다. * Narayana에 의하면 **User study task는 level of participant foresight에 의존적**일 수 있다. * **Incentivization (보상)** 역시 중요하다. |

[04. Taxonomy of Human Subject Evaluation in XAI – Study Design Dimensions]

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Evaluation에 대한 study design은 **질적일 수도 있고, 양적일 수도 있고, 아니면 혼합될** 수도 있다. (mixed study approach)   * 실험적인 연구에서 실험자는 실험 참여자 그룹에 treatment를 할당한다.   **<4 types of treatments combinations>**   |  |  | | --- | --- | | Single treatment | 대안적인 treatment가 없음 | | With and without explanation | **No explanation이 alternative treatment**일 수 있다. | | Alternative explanation | **예시:** treatment들과 user interface의 다른 aspect들 간 **설명에서 정보를 다양하게 제공** | | Alternative explanation interface | **예시:** treatment들 간에 user interface가 다양함 |   **<treatment assignment에 따른 study design의 구분>**   |  |  | | --- | --- | | Between-subjects designs | **실험 참여자 그룹 간의 이해의 차이**를 연구 | | Within-subject designs | **여러 개의 treatment**를 할당받은 **각 실험 참여자 간의 차이**를 연구 | |