# Ask Not What AI Can Do, But What AI Should Do: Towards a Framework of Task Delegability

Seungmin Lee (profile2697@gmail.com; 2013-11420), Dept. of Computer Science and Engineering, Seoul National University

## 1. Motivation of This Work

The recent development of artificial intelligence (AI) has made people look forward to the opportunities that the technology will bring to us. At the same time, however, there is growing concerns about which tasks and to what extent AI should be applied. These concerns can be naturally rewritten as a following question: Which tasks can be delegated to AI?

According to the paper, there is two dimensions that should be considered to answering the question. The first dimension is *what AI can do*, or *capability*. This means performance of AI for a particular task. Almost all traditional researches are about this dimension. The second dimension is *what AI should do*, or *human preference* which means what role human wants to delegate to AI. This is clearly important dimension, but virtually no research has been conducted. This work is the first empirical study about the second dimension and has been conducted for better understanding of how different factors affect human preferences of delegation (*delegability*).

## 2. Approach

For this work, the authors have done following three steps: devising a framework based on four factors that are considered to affect the delegability, constructing a dataset of diverse tasks like 'dignosing cancers' for understanding the delegabilities in different tasks, conducting a survey created using the factors and dataset to investigate to which extent people prefer AI's involvement.

### 2.1. Step 1: Devising a Framework for Delegability

**Four Factors** The authors devised a framework with four high-level factors considered to affect delegability: a person's **motivation**, task's **difficulty**, a person's subjective perception of **risk** that the one should take, and a person's **trust** in AI agent. These factors are further divided into sub-concepts as shown in Table 1.

**Delegability** The authors splited degree of delegation to measure human preferences according to the four factors. The degree of delegation is splited into the follwing four levels: **No AI assistance**, **The human leads and the AI assists** (or *machine-in-the-loop*), **The AI leads and the human assists** (or *human-in-the-loop*) and **Full AI automation**.

| Factors | Components |
|---|---|
| Motivation | Intrinsic motivation, goals, utility |
| Difficulty | Social skills, creativity, effort required, expertise required, human ability |
| Risk | Accountability, uncertainty, impact |
| Trust | Machine ability, interpretability, value alignment |

Table 1. An overview of the four factors in the framework.

### 2.2. Step 2: Constructing a Dataset of 100 Tasks

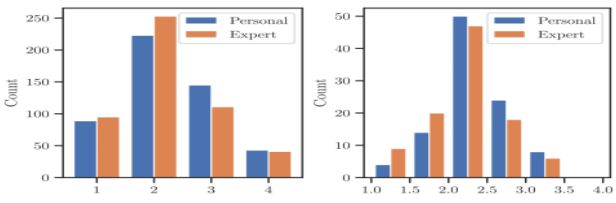
To evaluate the framework empirically, the authors collected diverse tasks from academic conferences, media, well-known occupations and people's everyday-life like 'Identifying fake news article'(from conferences) or 'buying a birthday present'(from everday-life).

### 2.3. Step 3: Conducting a Survey

The authors conducted 5-minute survey using Mechanical Turk. They used two versions of surveys: **Personal Survey** and **Expert Survey** because if a subject does not perform a task that requires expertise, e.g., 'diagnosing cancer', the motivation can not be estimated. They recorded total 1000 responses: 500 personal and expert surveys each and 5 for the 100 tasks each.

## 3. Results

On both personal and expert surveys, the subjects prefer *machine-in-the-loop*, and have little preference for *AI only* as we can see on Figure 1. Examining the correlation between the delegability and the four factors, trust showed the strongest correlation. However, interestingly, interpretability which belongs to sub-concepts of trust does not show significant relationships. They also investigated relations between factors and found difficulty and risk have the strongest correlation. Additionally, the authors trained a classifier for verfifying the proposed framework and that classifier shows better performance than a random baseline.




(a) Individual preferences. (b) Averaged preferences for each task.

Figure 1. Distributions of survey responses. 1 is Human only, and 4 is AI only