

Problem

- **Values** are abstract motivations that guide our opinions and actions.
- Engineering **value-sensitive agents** that learn and align their actions with human values is essential for robust and beneficial AI.
- **What values** should an agent elicit, learn, or align with?

Basic vs. Context-Specific Values

Basic values are:

- General and abstract;
- Applicable across contexts;
- Suitable for societal questions.

Context-specific values are:

- Applicable to a context;
- Defined within a context;
- Suitable for concrete usage.

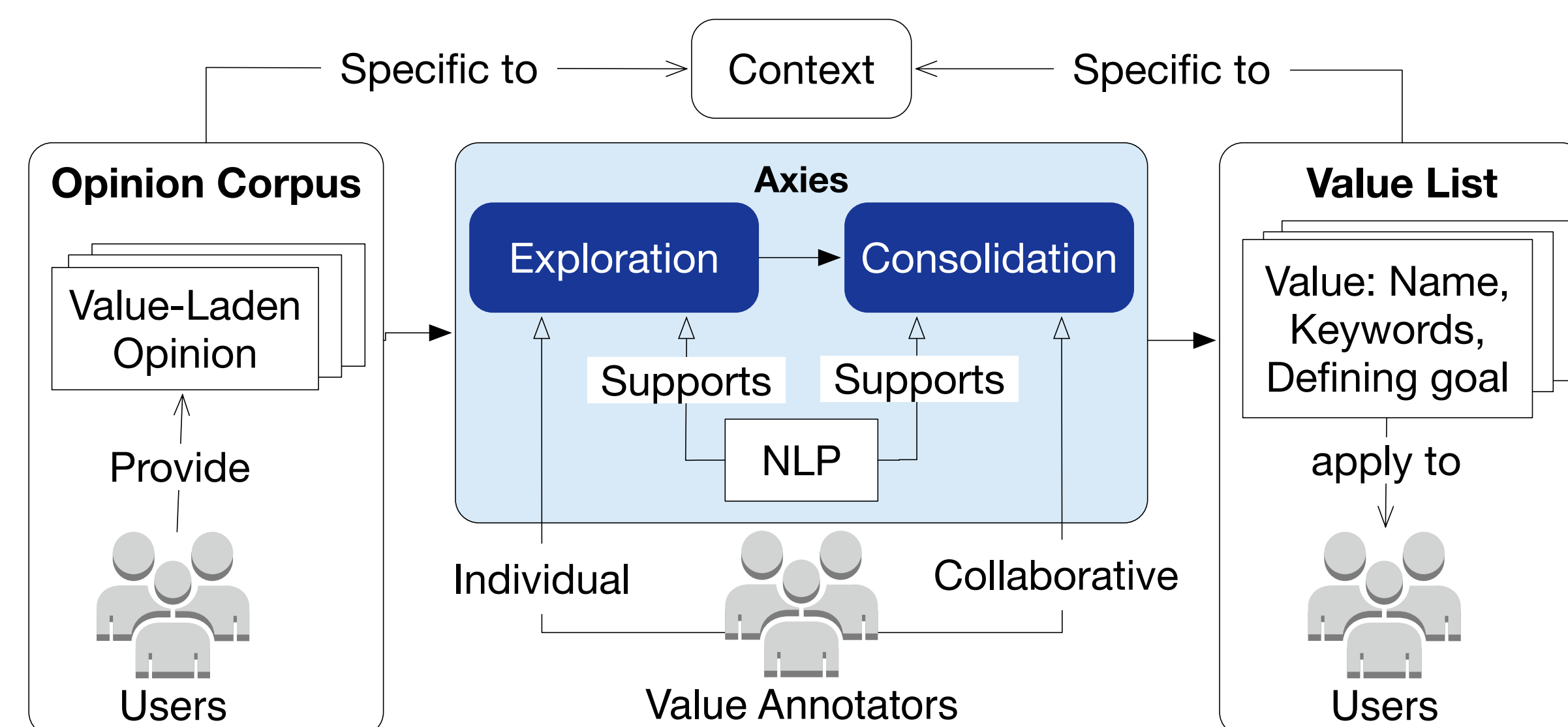
Contributions

As context-specific values vary with contexts, we need an efficient and reusable approach to identify them. We propose:

- **Axes**: a **methodology** for identifying context-specific values [1];
- A collaborative **web platform** to support Axes [2];
- An **evaluation** of Axes via a user study involving 60 subjects [1].

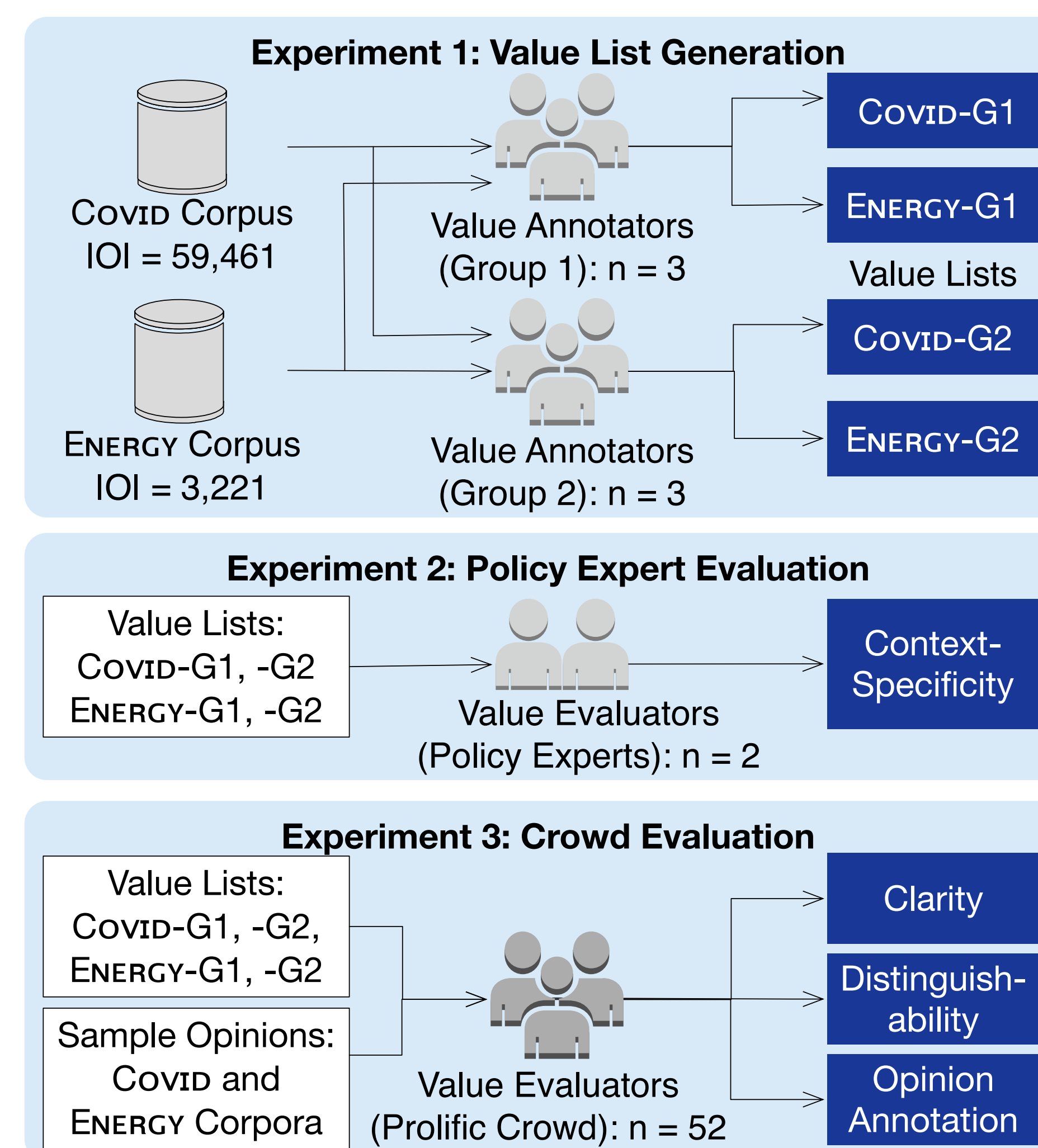
Axes Methodology

- Axes facilitates **inductive reasoning** and **collaborative work**;
- Axes exploits **natural language processing** and **active learning** techniques to guide annotation.

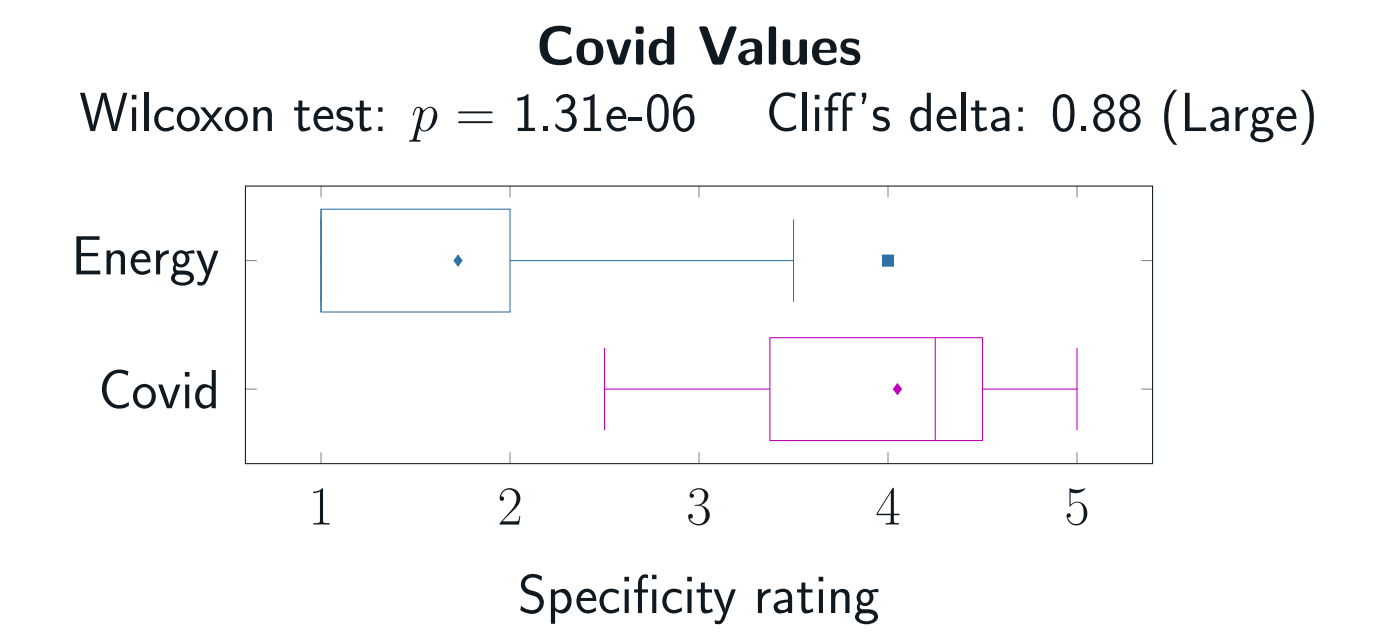


Web Platform

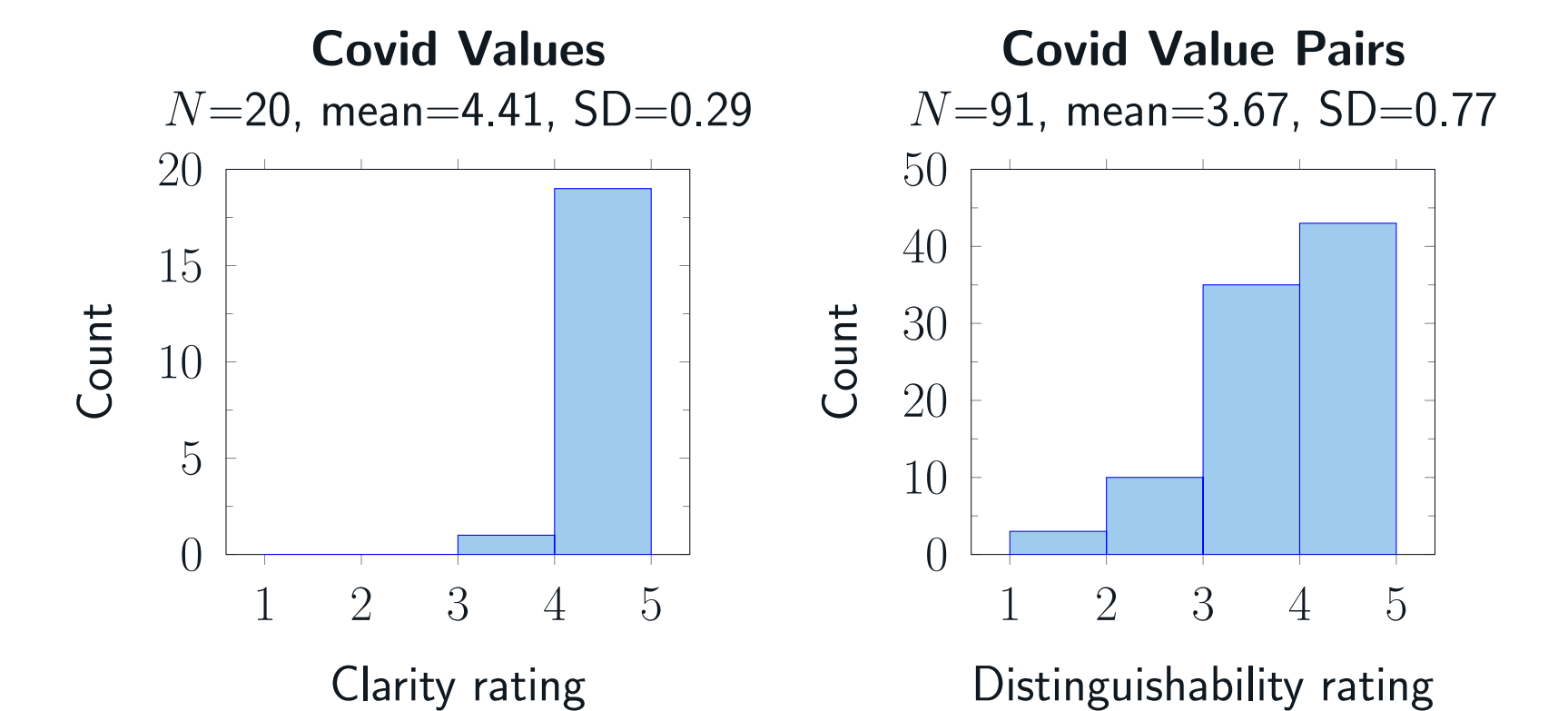
Experiments



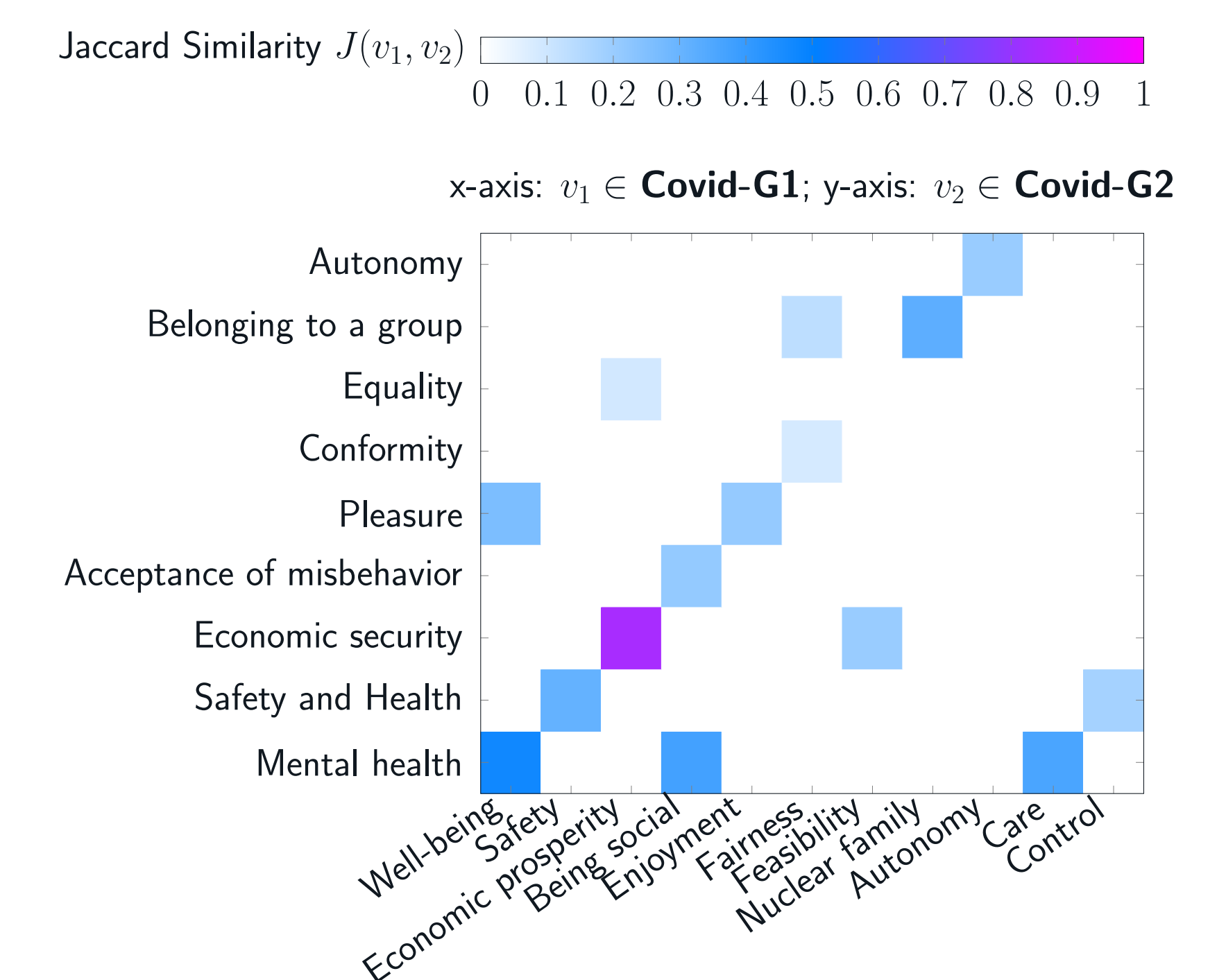
Axes Yields Context-Specific Values



Axes Yields Comprehensible Values



Axes Yields Consistent Value Lists



References

- [1] Enrico Liscio et al. "Axes: Identifying and Evaluating Context-Specific Values". In: *Proc. AAMAS 2021*, pp. 1–10.
- [2] Enrico Liscio et al. "A Collaborative Platform for Identifying Context-Specific Values: Demonstration Track". In: *Proc. AAMAS 2021*, pp. 1–3.