Axies: Identifying and Evaluating Context-Specific Values

E. Liscio,[1] M. van der Meer,[2] L. C. Siebert,[1] C. M. Jonker[1,2], N. Mouter,[1] and P. K. Murukannaiah[1]

[1]{E.Liscio, L.CavalcanteSiebert, C.M.Jonker, N. Mouter, P.K.Murukannaiah}@tudelft.nl [2]{M.T.van.der.Meer, C.M.Jonker}@liacs.leidenuniv.nl

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.

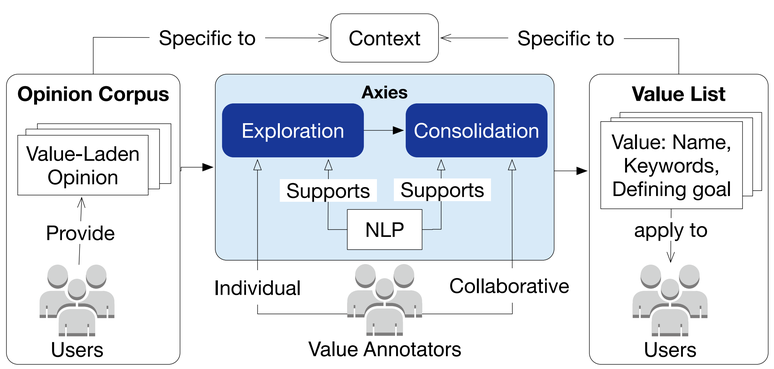
Contributions

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

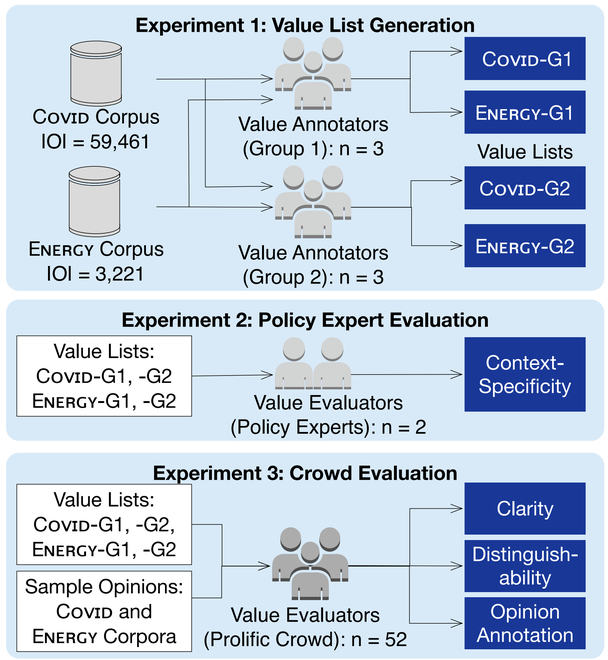
- Axies: a **methodology** for identifying context-specific values [1];
- A collaborative **web platform** to support Axies [2];
- An **evaluation** of Axies via a user study involving 60 subjects [1].

Axies Methodology

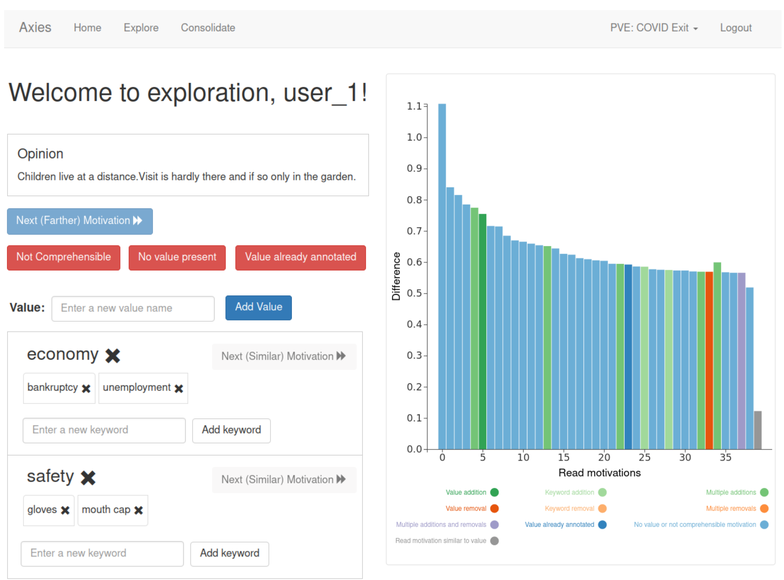
- Axies facilitates **inductive reasoning** and **collaborative work**;
- Axies exploits **natural language processing** and **active learning** techniques to guide annotation.



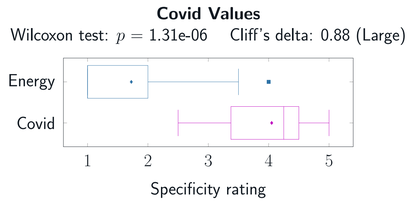
Experiments



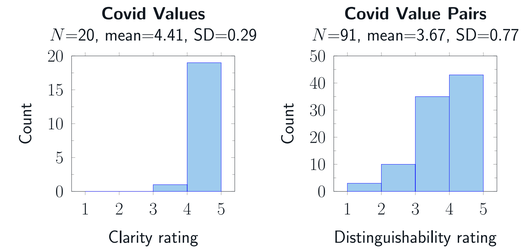
Web Platform



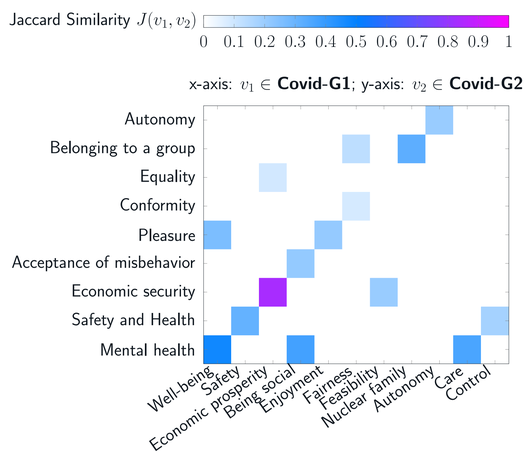
Axies Yields Context-Specific Values



Axies Yields Comprehensible Values



Axies Yields Consistent Value Lists



References

[1] Enrico Liscio et al. "Axies: Identifying and Evaluating Context-Specific Values". In: *Proc. AAMAS 2021*, pp. 799–808.

[2] Enrico Liscio et al. "A Collaborative Platform for Identifying Context-Specific Values: Demonstration Track". In: *Proc. AAMAS 2021*, pp. 1773–1775.