What Values should an Artificial Agent Align with?

Human values are the abstract motivations that justify our opinions and actions.

Engineering value-sensitive agents that align their actions with human values is essential for robust and beneficial AI.

General Values	Context-Specific Values
Broad and abstract	Applicable to a context
Applicable across contexts	Defined within a context
Suitable for societal questions	Suitable for concrete usage

Contributions

Axies: a methodology for identifying context-specific values

A collaborative web platform to support Axies

An evaluation of Axies via a user study involving 80 subjects

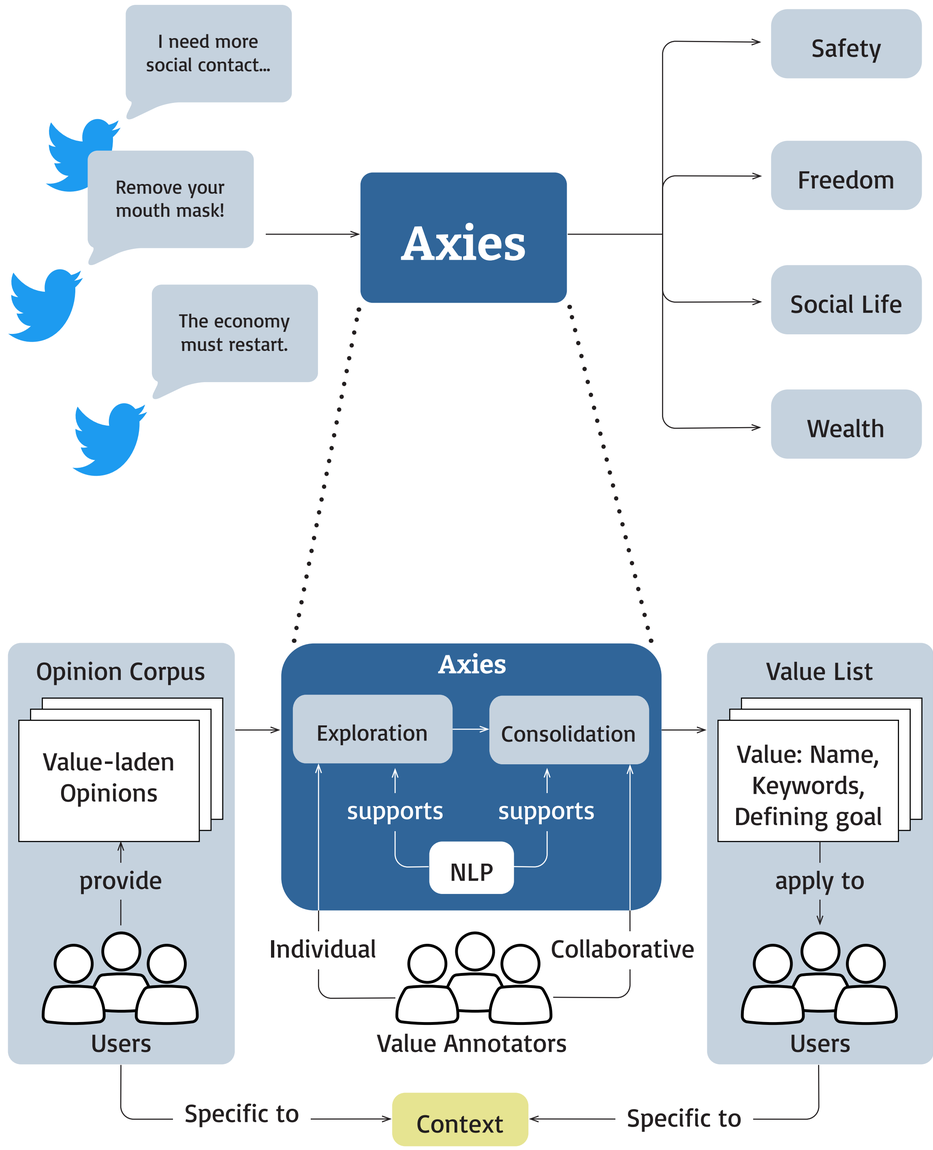
A comparison of general and context-specific values

Enrico Liscio, Michiel van der Meer, Luciano C. Siebert, Catholijn M. Jonker, Pradeep K. Murukannaiah

TU Delft

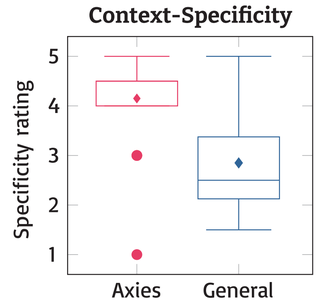
Hybrid Intelligence

A Hybrid Methodology for Identifying Context-Specific Values

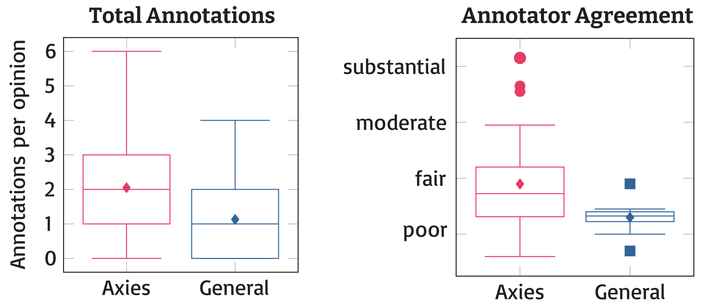


Context-specific values are more suited for concrete applications than general values.

Axies values are more context-specific than general values.



Axies values are annotated more often and with higher annotator agreement than general values.



Not all general values are relevant to a context; there is a one-to-many relationship between general and Axies values.