A Review of AI Agent Reasoning with Values

Jay Paul Morgan & Adam Wyner

Swansea University
Prifysgol Abertawe

Computational Foundry
Ffowndri Gyfrifiadol

25th October 2025
Value Engineering in AI (VALE) 2025

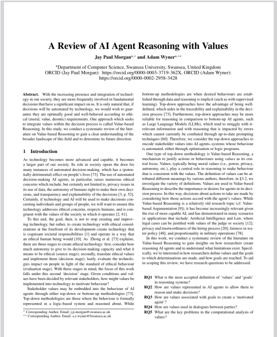
Introduction

What we have done:

- Performed a systematic review of the literature on Value-based Reasoning.
- Articles for consideration were selected using a standardised (systematic) process.
- From the 57 articles in question, information was extracted to answer 5 research questions.

In this talk:

1. A brief description of the systematic framework.
2. 3 of the research questions and what has been found.
3. Conclusions & future direction.



A Review of AI Agent Reasoning with Values

Jay Paul Morgan[a,*] and Adam Wyner[a,**]

[a]Department of Computer Science, Swansea University, Swansea, United Kingdom
ORCID (Jay Paul Morgan): https://orcid.org/0000-0003-3719-362X, ORCID (Adam Wyner):
https://orcid.org/0000-0002-2958-3428

Abstract. With the increasing presence and integration of technology in our society, they are more frequently involved in fundamental decisions that have a significant impact on us. It is only natural that, if decisions will be automated by technology, we would wish to guarantee they are optimally good and well-behaved according to ethical (moral, value, deontic) requirements. One approach which seeks to integrate values within the decision process is called Value-based Reasoning. In this study, we conduct a systematic review of the literature on Value-based Reasoning to gain a clear understanding of the broader landscape of this field and to determine its future direction.

1 Introduction

As technology becomes more advanced and capable, it becomes a larger part of our society. Its role in society opens the door for many instances of automated decision-making, which has a (potentially detrimental) effect on people's lives [73]. The use of automated decision-making AI agents, in particular, raises numerous ethical concerns which include, but certainly not limited to, privacy issues in its use of data, the autonomy of humans right to make their own decisions, and transparency and explainability of the decisions [3, p. 52]. Certainly, if technology and AI will be used to make decisions concerning individuals and groups of people, we will want to ensure this technology addresses ethical concerns, respects humans, and is congruent with the values of the society in which it operates [2, 41].

To this end, the goal, then, is not to stop creating and improving technology, but rather, to create technology with ethical considerations at the forefront of its development–create technology that is cognisant societal responsibilities [1] and operate in a way that an ethical human being would [10]. As Zhong et al. [73] explains, there are three stages to create ethical technology: first, consider how much autonomy to give to its decision-making capacity and what it means to be ethical (source stage); secondly, translate ethical values and implement them (decision stage); lastly, evaluate the technologies impact on people in light of the standard of ethical behaviour (evaluation stage). With these stages in mind, the focus of this work falls under this second 'decision' stage. Given conditions and values have been decided by relevant stakeholders, how might values be implemented into technology to motivate behaviour?

Stakeholder values may be embedded into the behaviour of AI agents through either top-down or bottom-up methodologies [73]. Top-down methodologies are those where the behaviour is formally represented in a logic-based system and reasoned about. While bottom-up methodologies are when desired behaviours are established through data and reasoning is implicit (such as with supervised learning). Top-down approaches have the advantage of being well-defined, which aides in the traceability and explainability in the decision process [73]. Furthermore, top-down approaches may be more reliable for reasoning in comparison to bottom-up AI agents, such as Large Language Models (LLMs), which tend to struggle with irrelevant information and with reasoning that is impacted by errors which cannot currently be combated through up-to-date prompting techniques [60]. Therefore, we consider the top-down approaches to encode stakeholder values into AI agents–systems whose behaviour is automated, either through optimisation or logic programs.

One type of top-down methodology is Value-based Reasoning, a mechanism to justify actions or behaviours using *values* as its central focus. Values, typically being moral values (i.e., power, privacy, autonomy, etc.), play a central role in reasoning to make behaviour that is consistent with the values. The definition of values can be attributed different meanings by various authors, therefore, in §3.2, we investigate the variety of definitions. Values are used in Value-based Reasoning to describe the importance or desires for agents in its decision process. In this way, decisions about actions to take are made by considering how those actions accord with the agent's values. While Value-based Reasoning is a relatively old research topic (cf. Value-based Argumentation [9]), it has become increasingly relevant given the rise of more capable AI, and has demonstrated in many scenarios or applications that include: Artificial Intelligence and Law, where legal cases can be justified with values of the judge or society [12]; privacy and trustworthiness of the hiring process [20]; fairness in water policy [48]; and proportionality in military operations [78].

In this work, we conduct a systematic review of the literature on Value-based Reasoning to gain insights on how researchers create reasoning AI agents and to understand what limitations exist. Specifically, we're interested in how researchers define values and the goals to which determinations are made, and how goals are reached. To aid in scoping this review, we have research questions to be addressed:

RQ1 What is the most accepted definition of 'values' and 'goals' in reasoning systems?
RQ2 How are values represented in AI agents to allow them to reason and make decisions?
RQ3 How are values associated with goals to create a 'motivated agent'?
RQ4 How are values used in dialogues between parties?
RQ5 What are the key problems in the computational analysis of values?

[*] Corresponding Author. Email: j.p.morgan@swansea.ac.uk
[**] Corresponding Author. Email: a.z.wyner@Swansea.ac.uk

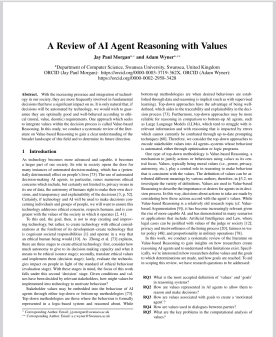
Introduction

What we have done:

- Performed a systematic review of the literature on Value-based Reasoning.
- Articles for consideration were selected using a standardised (systematic) process.
- From the 57 articles in question, information was extracted to answer 5 research questions.

In this talk:

1. A brief description of the systematic framework.
2. 3 of the research questions and what has been found.
3. Conclusions & future direction.



A Review of AI Agent Reasoning with Values

Jay Paul Morgan[a,*] and Adam Wyner[a,**]

[a]Department of Computer Science, Swansea University, Swansea, United Kingdom
ORCID (Jay Paul Morgan): https://orcid.org/0000-0003-3719-362X, ORCID (Adam Wyner): https://orcid.org/0000-0002-2958-3428

Abstract. With the increasing presence and integration of technology in our society, they are more frequently involved in fundamental decisions that have a significant impact on us. It is only natural that, if decisions will be automated by technology, we would wish to guarantee they are optimally good and well-behaved according to ethical (moral, value, deontic) requirements. One approach which seeks to integrate values within the decision process is called Value-based Reasoning. In this study, we conduct a systematic review of the literature on Value-based Reasoning to gain a clear understanding of the broader landscape of this field and to determine its future direction.

1 Introduction

As technology becomes more advanced and capable, it becomes a larger part of our society. Its role in society opens the door for many instances of automated decision-making, which has a (potentially detrimental) effect on people's lives [73]. The use of automated decision-making AI agents, in particular, raises numerous ethical concerns which include, but certainly not limited to, privacy issues in its use of data, the autonomy of humans right to make their own decisions, and transparency and explainability of the decisions [3, p. 52]. Certainly, if technology and AI will be used to make decisions concerning individuals and groups of people, we will want to ensure this technology addresses ethical concerns, respects humans, and is congruent with the values of the society in which it operates [2, 41].

To this end, the goal, then, is not to stop creating and improving technology, but rather, to create technology with ethical considerations at the forefront of its development–create technology that is cognisant societal responsibilities [1] and operate in a way that an ethical human being would [10]. As Zhong et al. [73] explains, there are three stages to create ethical technology: first, consider how much autonomy to give to its decision-making capacity and what it means to be ethical (source stage); secondly, translate ethical values and implement them (decision stage); lastly, evaluate the technologies impact on people in light of the standard of ethical behaviour (evaluation stage). With these stages in mind, the focus of this work falls under this second 'decision' stage. Given conditions and values have been decided by relevant stakeholders, how might values be implemented into technology to motivate behaviour?

Stakeholder values may be embedded into the behaviour of AI agents through either top-down or bottom-up methodologies [73]. Top-down methodologies are those where the behaviour is formally represented in a logic-based system and reasoned about. While bottom-up methodologies are when desired behaviours are established through data and reasoning is implicit (such as with supervised learning). Top-down approaches have the advantage of being well-defined, which aides in the traceability and explainability in the decision process [73]. Furthermore, top-down approaches may be more reliable for reasoning in comparison to bottom-up AI agents, such as Large Language Models (LLMs), which tend to struggle with irrelevant information and with reasoning that is impacted by errors which cannot currently be combated through up-to-date prompting techniques [60]. Therefore, we consider the top-down approaches to encode stakeholder values into AI agents–systems whose behaviour is automated, either through optimisation or logic programs.

One type of top-down methodology is Value-based Reasoning, a mechanism to justify actions or behaviours using *values* as its central focus. Values, typically being moral values (i.e., power, privacy, autonomy, etc.), play a central role in reasoning to make behaviour that is consistent with the values. The definition of values can be attributed different meanings by various authors, therefore, in §3.2, we investigate the variety of definitions. Values are used in Value-based Reasoning to describe the importance or desires for agents in its decision process. In this way, decisions about actions to take are made by considering how those actions accord with the agent's values. While Value-based Reasoning is a relatively old research topic (cf. Value-based Argumentation [9]), it has become increasingly relevant given the rise of more capable AI, and has demonstrated in many scenarios or applications that include: Artificial Intelligence and Law, where legal cases can be justified with values of the judge or society [12]; privacy and trustworthiness of the hiring process [20]; fairness in water policy [48]; and proportionality in military operations [78].

In this work, we conduct a systematic review of the literature on Value-based Reasoning to gain insights on how researchers create reasoning AI agents and to understand what limitations exist. Specifically, we're interested in how researchers define values and the goals to which determinations are made, and how goals are reached. To aid in scoping this review, we have research questions to be addressed:

RQ1 What is the most accepted definition of 'values' and 'goals' in reasoning systems?
RQ2 How are values represented in AI agents to allow them to reason and make decisions?
RQ3 How are values associated with goals to create a 'motivated agent'?
RQ4 How are values used in dialogues between parties?
RQ5 What are the key problems in the computational analysis of values?

[*] Corresponding Author. Email: j.p.morgan@swansea.ac.uk
[**] Corresponding Author. Email: a.z.wyner@Swansea.ac.uk

Search Methodology

Table: PICO keywords and synonyms used to search digital libraries.

	Keywords	Synonyms
Population	AI	AI Agents, Artificial Intelligence, Machine Learning, Multi-Agent Systems
Intervention	Value-based Reasoning	Value-based argumentation
Comparison	Logic, Argumentation, Reasoning, Philosophy, Value, Computational Model	Norm
Outcome	Representation, Dialogue, Behaviour	Negotiation, Persuasion, Action

Search Methodology

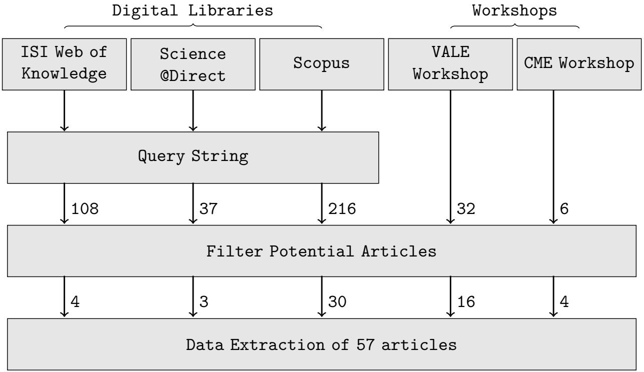
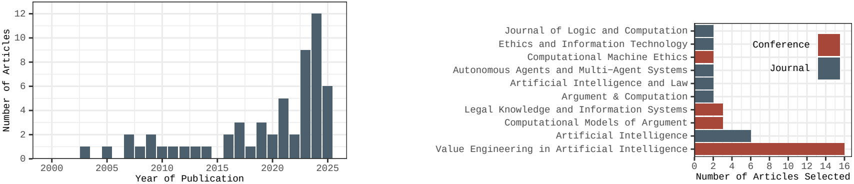


Figure: Overview of the article selection process.

Search Methodology



(a) The number of articles selected from 2000-2025.

(b) The top-10 sources from which articles were selected.

Research Questions

RQ1

What is the most accepted definition of ‘values’ and ‘goals’ in reasoning systems?

RQ2

How are values represented in AI agents to allow them to reason and make decisions?

RQ3

How are values associated with goals to create a ‘motivated agent’?

RQ4

How are values used in dialogues between parties?

RQ5

What are the key problems in the computational analysis of values?

RQ1: What is the most accepted definition of ‘values’ and ‘goals’ in reasoning systems?

How do different authors define ‘values’ and ‘goals’ and do these have an effect on the formalisations? And is there a consensus?

Values:

- Values are “Abstract principles that guide behaviour”
- Many look to the Schwartz’s Theory of Basic Human Values (STBHV).
- When Schwartz is used, sometimes only a some of values are used–which limit the usefulness of STBHV.

Goals:

- Tend to be less defined.
- “Goals reflect the state of affairs the agent wishes to bring about”.
- Some types of Goals:
    - Achievement (Make false → true)
    - Remedy (Make true → false)
    - Maintenance (Keep true, true)
    - Avoidance (Keep false, false)

Research Questions

RQ1

What is the most accepted definition of ‘values’ and ‘goals’ in reasoning systems?

RQ2

How are values represented in AI agents to allow them to reason and make decisions?

RQ3

How are values associated with goals to create a ‘motivated agent’?

RQ4

How are values used in dialogues between parties?

RQ5

What are the key problems in the computational analysis of values?

RQ2: How are values represented in AI agents to allow them to reason and make decisions?

Throughout the articles, we see two methodologies for creating such reasoning systems:

(1) through value-alignment, meaning values are implicitly encoded into the system through its output behaviour;

(2) an explicit representation of values where determinations of inner behaviour and action are reasoned through states/functions that represent the values of interest.

Implicit

- No explicit representation.
- System performs instrumental actions (actions that are instrumental to bring out goals which align with values).

Explicit

- State/Object
  $V = \{v_1, v_2, \ldots, v_n\}$
- Function
  $f(\text{action})_v \rightarrow \{+, -, =\}$
- Numerical
  $v_1 = 0.25, v_2 = 0.5, \ldots$

Research Questions

RQ1

What is the most accepted definition of ‘values’ and ‘goals’ in reasoning systems?

RQ2

How are values represented in AI agents to allow them to reason and make decisions?

RQ3

How are values associated with goals to create a ‘motivated agent’?

RQ4

How are values used in dialogues between parties?

RQ5

What are the key problems in the computational analysis of values?

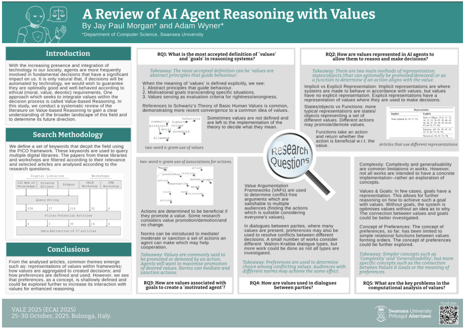
RQ5: What are the key problems in the computational analysis of values?

Lastly, we evaluate the key issues with the formalisations that might point to the future direction of the field.

- **Deeper and more realistic scenarios** Work should be done to create datasets or scenarios with which to compare methodologies.
- Connection between **Values & Goals**. While 'value' has had more attention in definition, the definition of 'goal' remains somewhat implicit (with only a few articles giving some definition and types). Furthermore, the connection between short-term and longer-term goals and how these are strategised with values can be explored in more depth (**Planning**).
- Concept of **Preferences**. Preferences have been limited to ordering relations, but this concept could be taken a lot further.

Conclusion

- Conducted a literature review that has identified main themes such as a general direction on the meaning of Values.
- Some concepts can be taken further: Connection between Values & Goals; Preferences.
- This presentation has only touched upon the main points, but there is more. Come and ask some questions at the poster session!



A Review of AI Agent Reasoning with Values

By Jay Paul Morgan* and Adam Wyner*

*Department of Computer Science, Swansea University

Introduction

With the increasing presence and integration of technology in our society, agents are more frequently involved in fundamental decisions that have a significant impact on us. It is only natural that, if decisions will be automated by technology, we would wish to guarantee they are optimally good and well-behaved according to ethical (moral, value, deontic) requirements. One approach which seeks to integrate values within the decision process is called Value-based Reasoning. In this study, we conduct a systematic review of the literature on Value-based Reasoning to gain a clear understanding of the broader landscape of this field and to determine its future direction.

Search Methodology

We define a set of keywords that decpit the field using the PICO framework. These keywords are used to query multiple digital libraires. The papers from these libraries and workshops are filtered according to their relevance and selected articles are analysed according to the research questions.

Digital Libraries: ISI Web of Knowledge, Science @Direct, Scopus → Query String → 108, 37, 216
Workshops: VALE Workshop, CMF Workshop → 32, 6
→ Filter Potential Articles → 4, 3, 30, 16, 4
→ Data Extraction of 57 articles

Conclusions

From the analysed articles, common themes emerge such as: representations of values within frameworks; how values are aggregated to created decisions; and how preferences are defined and used. However, we see that preferences, as a concept, is shallowly defined and could be explored further to increase its interaction with values for enhanced reasoning.

RQ1: What is the most accepted definition of `values' and `goals' in reasoning systems?

Takeaway: The most accepted definition can be 'values are abstract principles that guide behaviour'.

When the meaning of 'values' is defined explicitly, we see:
1. Abstract principles that guide behaviour.
2. Motivational goals transcending specific situations.
3. Values serving as evaluation criteria for rightness/wrongness.

References to Schwartz's Theory of Basic Human Values is common, demonstrating more recent convergence to a common idea of values.

two-word n-gram use of values.

Sometimes values are not defined and are left to the implementation of the theory to decide what they mean.

RQ2: How are values represented in AI agents to allow them to reason and make decisions?

Takeaway: There are two main methods of representation: states/objects (that can optionally be promoted/demoted) or as a function to determine if an action aligns with the value.

Implicit vs Explicit Representation: Implicit representations are where systems are made to behave in accordance with values, but values have no explicit representation. Explicit representations have a representation of values where they are used to make decisions.

States/objects vs Functions: more typical representations are states/ objects representing a set of different values. Different actions may promote/demote values.

Functions take an action and return whether the action is beneficial w.r.t. the value.

Representation	
Implicit	Explicit
Value-aligned [4, 16, 17, 21]	State or Object: [5–8, 11, 13, 14, 22, 23, 26–35, 39, 40, 43, 44, 46, 47, 49, 58, 59, 63–65, 67, 69–72, 77, 79]
	Function: [18, 26, 38, 45, 47, 52, 57, 62, 74–76, 78]
	Numerical: [36, 42, 74, 75]

Articles that use different representations

Research Questions

two-word n-gram use of associations for actions.

Actions are determined to be beneficial if they promote a value. Some research considers value promotion/demotion/and no change.

Norms can be introduced to mediate/ moderate or sanction a set of actions an agent can make which may help cooperation.

Takeaway: Values are commonly said to be promoted or demoted by an action. Agents will want to maximise promotion of desired values. Norms can mediate and sanction actions.

RQ3: How are values associated with goals to create a 'motivated agent'?

Value Argumentation Frameworks (VAFs) are used to determine conflict-free arguments which are satisfiable to multiple audiences (finding the actions which is suitable considering everyone's values).

In dialogues between parties, where many values are present, preferences may also be used to resolve conflicts between different decisions. A small number of works consider different Walton-Krabbe dialogue types, but more work could be done as not all types are investigated.

Takeaway: Preferences are used to determine choice among conflicting values. Audiences with diffferent norms may achieve the same effect.

RQ4: How are values used in dialogues between parties?

Complexity: Complexity and generalisability are common limitations in works. However, not all works are intended to have a concrete implementation--rather an exploration of concepts.

Values & Goals: In few cases, goals have a representation. This allows for further reasoning on how to achieve such a goal with values. Without goals, the system is optimises values without an idea as to why. The connection between values and goals could be better investigated.

Concept of Preferences: The concept of preferences, so far, has been limited to simple relational functions between values forming orders. The concept of preferences could be further explored.

Takeaway: Simpler concepts such as 'Complexity' and 'Generalisability', but more specific concepts such as the connection between Values & Goals or the meaning of preferences.

RQ5: What are the key problems in the computational analysis of values?

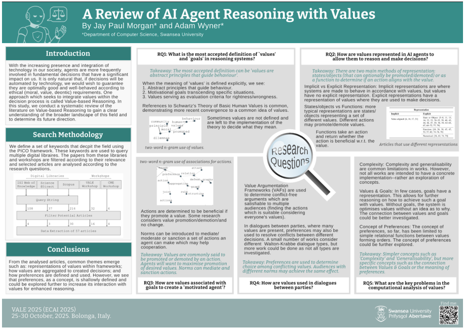
VALE 2025 (ECAI 2025)
25-30 October, 2025. Bolonga, Italy.

Swansea University Prifysgol Abertawe

Find me:

Conclusion

- Conducted a literature review that has identified main themes such as a general direction on the meaning of Values.
- Some concepts can be taken further: Connection between Values & Goals; Preferences.
- This presentation has only touched upon the main points, but there is more. Come and ask some questions at the poster session!



A Review of AI Agent Reasoning with Values

By Jay Paul Morgan* and Adam Wyner*

*Department of Computer Science, Swansea University

Introduction

With the increasing presence and integration of technology in our society, agents are more frequently involved in fundamental decisions that have a significant impact on us. It is only natural that, if decisions will be automated by technology, we would wish to guarantee they are optimally good and well-behaved according to ethical (moral, value, deontic) requirements. One approach which seeks to integrate values within the decision process is called Value-based Reasoning. In this study, we conduct a systematic review of the literature on Value-based Reasoning to gain a clear understanding of the broader landscape of this field and to determine its future direction.

Search Methodology

We define a set of keywords that decpit the field using the PICO framework. These keywords are used to query multiple digital libraires. The papers from these libraries and workshops are filtered according to their relevance and selected articles are analysed according to the research questions.

Digital Libraries: ISI Web of Knowledge, Science @Direct, Scopus → Query String → 108, 37, 216

Workshops: VALE Workshop → 32; CME Workshop → 6

Filter Potential Articles → 4, 3, 30, 16, 4

Data Extraction of 57 articles

Conclusions

From the analysed articles, common themes emerge such as: representations of values within frameworks; how values are aggregated to created decisions; and how preferences are defined and used. However, we see that preferences, as a concept, is shallowly defined and could be explored further to increase its interaction with values for enhanced reasoning.

RQ1: What is the most accepted definition of `values' and `goals' in reasoning systems?

Takeaway: The most accepted definition can be 'values are abstract principles that guide behaviour'.

When the meaning of 'values' is defined explicitly, we see:
1. Abstract principles that guide behaviour.
2. Motivational goals transcending specific situations.
3. Values serving as evaluation criteria for rightness/wrongness.

References to Schwartz's Theory of Basic Human Values is common, demonstrating more recent convergence to a common idea of values.

two-word n-gram use of values.

Sometimes values are not defined and are left to the implementation of the theory to decide what they mean.

RQ2: How are values represented in AI agents to allow them to reason and make decisions?

Takeaway: There are two main methods of representation: states/objects (that can optionally be promoted/demoted) or as a function to determine if an action aligns with the value.

Implicit vs Explicit Representation: Implicit representations are where systems are made to behave in accordance with values, but values have no explicit representation. Explicit representations have a representation of values where they are used to make decisions.

States/objects vs Functions: more typical representations are states/objects representing a set of different values. Different actions may promote/demote values.

Functions take an action and return whether the action is beneficial w.r.t. the value.

Representation	
Implicit	Explicit
Value-aligned [4, 16, 17, 21]	State or Object: [5–8, 11, 13, 14, 22, 23, 26–35, 39, 40, 43, 44, 46, 47, 49, 58, 59, 63–65, 67, 69–72, 77, 79]
	Function: [18, 26, 38, 45, 47, 52, 57, 62, 74–76, 78]
	Numerical: [36, 42, 74, 75]

Articles that use different representations

Research Questions

two-word n-gram use of associations for actions.

Actions are determined to be beneficial if they promote a value. Some research considers value promotion/demotion/and no change.

Norms can be introduced to mediate/moderate or sanction a set of actions an agent can make which may help cooperation.

Takeaway: Values are commonly said to be promoted or demoted by an action. Agents will want to maximise promotion of desired values. Norms can mediate and sanction actions.

RQ3: How are values associated with goals to create a 'motivated agent'?

Value Argumentation Frameworks (VAFs) are used to determine conflict-free arguments which are satisfiable to multiple audiences (finding the actions which is suitable considering everyone's values).

In dialogues between parties, where many values are present, preferences may also be used to resolve conflicts between different decisions. A small number of works consider different Walton-Krabbe dialogue types, but more work could be done as not all types are investigated.

Takeaway: Preferences are used to determine choice among conflicting values. Audiences with diffferent norms may achieve the same effect.

RQ4: How are values used in dialogues between parties?

Complexity: Complexity and generalisability are common limitations in works. However, not all works are intended to have a concrete implementation--rather an exploration of concepts.

Values & Goals: In few cases, goals have a representation. This allows for further reasoning on how to achieve such a goal with values. Without goals, the system is optimises values without an idea as to why. The connection between values and goals could be better investigated.

Concept of Preferences: The concept of preferences, so far, has been limited to simple relational functions between values forming orders. The concept of preferences could be further explored.

Takeaway: Simpler concepts such as 'Complexity' and 'Generalisability', but more specific concepts such as the connection between Values & Goals or the meaning of preferences.

RQ5: What are the key problems in the computational analysis of values?

VALE 2025 (ECAI 2025)
25-30 October, 2025. Bolonga, Italy.

Swansea University
Prifysgol Abertawe

Find me: