Understanding the Process of Human-AI Value Alignment

JACK MCKINLAY*, University of Bath, UK
MARINA DE VOS, University of Bath, UK
JANINA A. HOFFMANN, University of Bath, UK
ANDREAS THEODOROU

Background: Value alignment in computer science research is often used to refer to the process of aligning the behaviour of artificial intelligence systems with humans' desires, but the way the phrase is used often lacks precision.
Objectives: In this paper, we conduct a systematic literature review to advance the understanding of value alignment in artificial intelligence by characterising the topic in the context of its research literature. We use this to suggest a more precise definition of the term.
Methods: We analyse the abstracts, introductions and conclusions of 172 value alignment research articles that have been published in recent years and synthesise their content using thematic analysis. From these 172 papers we select 85 papers using a structured criteria for a deep analysis, coding these papers in full.
Results: Our analysis leads to six themes: value alignment drivers & approaches; challenges in value alignment; values in value alignment; cognitive processes in humans and AI; human-agent teaming; and designing and developing value-aligned systems.
Conclusions: By analysing these themes in the context of the literature, we define value alignment as an ongoing process between humans and autonomous agents that aims to express and implement abstract values in diverse contexts, while managing the cognitive limits of both humans and AI agents and also balancing the conflicting ethical and political demands generated by the values in different groups. Our analysis gives rise to a set of research challenges and opportunities in the field of value alignment for future work.

JAIR Track: Surveys

JAIR Associate Editor: Paolo Turrini

JAIR Reference Format:
Jack McKinlay, Marina De Vos, Janina A. Hoffmann, and Andreas Theodorou. 2026. Understanding the Process of Human-AI Value Alignment. *Journal of Artificial Intelligence Research* 85, Article 29 (March 2026), 41 pages. DOI: 10.1613/jair.1.18846

1 Introduction

Value alignment is broadly understood as the challenge of ensuring that autonomous artificial agents act in ways aligned with humans and their values when deployed in society (Russell 2019, p.137). However, this is complicated by the complexity and diversity of human values, and their abstract, generalisable nature.

There are hundreds of guidelines for how artificial intelligence (AI) systems should be developed and deployed, but values are too often described using non-specific language (Aler Tubella, Theodorou, F. Dignum, et al. 2019; Theodorou and V. Dignum 2020). This requires developers to interpret values in the problem context during

*Corresponding Author.

Authors' Contact Information: Jack McKinlay, ORCID: 0000-0001-9822-8166, jam218@bath.ac.uk, University of Bath, Bath, UK; Marina De Vos, ORCID: 0000-0003-3583-7671, cssmdv@bath.ac.uk, University of Bath, Bath, UK; Janina A. Hoffmann, ORCID: 0000-0002-6246-2724, jah253@bath.ac.uk, University of Bath, Bath, UK; Andreas Theodorou, ORCID: 0000-0001-9499-1535, a.theodorou@bath.edu.

DOI: 10.1613/jair.1.18846

system development. However, the subjective nature of this interpretation may result in inconsistencies in how values are achieved in different contexts. This subjectivity is further complicated by the opaque nature of many AI systems, which makes determining the influence of values in systems and their impacts difficult (Poel 2020; Umbrello and Poel 2021).

Values can be described as the drivers in individual and shared decision-making in humans, as guiding factors shared across cultures (Schwartz and Bilsky 1987). A mutual understanding of each others' values enables us to cooperate on tasks despite differing backgrounds and interests. As autonomous agents become more prevalent as actors in society, their actions may not only realize human values but also shape the development of these values (M. Cappuccio et al. 2021). To fully benefit from AI as a technology, it is essential to ensure that these autonomous agents act in accordance with our values so that their actions lead to desirable outcomes.

While researchers generally agree on the same high-level understanding of value alignment, definitions of value alignment presented in the literature are diverse and often shallow, if they are specified in the paper at all. Some authors take the view that value alignment should focus on coordinating objective functions (Sanneman and Shah 2023b), which likely stems from the influential paper on cooperative inverse reinforcement learning by Hadfield-Menell, Dragan, et al. (2016). Other authors like Fisac et al. (2020), Lera-Leri et al. (2022), and Vamplew et al. (2018) provide a more high-level definition, which gives a broad goal but lacks direction in achieving it. Serramia, Lopez-Sanchez, et al. (2020) and Serramia, Rodriguez-Soto, et al. (2023) and Sierra et al. (2021) provide definitions that are more focused on alignment between norms and values but are still explicitly referred to as value alignment. Given these diverse interpretations, papers with missing definitions (Bogosian 2017; Peschl et al. 2022; Siebert et al. 2022) are problematic, as it makes it ambiguous how the authors have interpreted the problem of value alignment and how their work relates to it.

This inconsistent terminology brought on by missing and conflicting definitions is understandable given the diversity of value alignment research. However, it still creates significant barriers for researchers by making it harder to develop a shared interpretation of the problem, and by posing challenges for researchers trying to assess the state of the value alignment field and identify useful research directions. The goal of this research is to provide a shared interpretation of the value alignment problem by analysing the different themes of value alignment research and to develop a conceptual understanding of value alignment as a process.

In order to develop our understanding of what value alignment is, we conduct a structured review of value alignment literature and analyse published articles using inductive thematic analysis. We take an inductive approach as this is suitable when no prior theory to explain a concept has emerged (Naeem et al. 2023). It also enables us to build broader themes from participants' views (Braun and Clarke 2006) while remaining open to new themes that may emerge in the analysis (Fereday and Muir-Cochrane 2006).

Our contributions in this paper are:

- a thematic analysis of 172 value alignment papers for the sake of extracting and analysing the core themes of value alignment research;
- an in-depth analysis of the concepts and challenges in value alignment based on the identified themes;
- identification of valuable research directions within these themes based on the existing literature.

The rest of the paper proceeds as follows:

- In Section 2 we discuss related works analysing value alignment literature and compare their approaches with our own.
- In Section 3 we explain our survey methodology, discussing our literature sources and our selection criteria. We also detail the number and types of papers returned in our search.
- Section 4 contains our analysis of our search results, and brings them together into our thematic analysis of the value alignment literature.

- In Section 5 we discuss the results of our survey and analysis, propose directions for future research in the field, and consider the limitations of our study.
- Finally, Section 6 concludes our analysis by reiterating the key observations from our analysis and their implications for value alignment work going forwards.

2 Related Works

Wallach and Allen (2008) is one of the earliest reviews on embedding human values in artificial intelligence. The review provides a thorough introduction to the problem of learning values and distinguishes between top-down, bottom-up and hybrid ethical learning. Although technology has naturally developed since the book's publication, the ideas were cited frequently throughout our review and hence can be expected to underpin the outcomes of our own analysis.

Tolmeijer et al. (2021) conducted a thorough review of the field of machine ethics, which value alignment has a close relationship with. This paper focused on machine ethics implementation in particular. The authors produced a trimorphic taxonomy of the literature based on: i) the ethical theory included; ii) non-technical aspects; and iii) technical aspects. While the work contributes a means to categorise and analyse broad aspects of the field using this taxonomy, our approach instead focuses on identifying themes found across papers to produce a practical understanding of value alignment as a process without focusing on individual ethical theories. This leads to our synthesis of literature into a conceptual process, a distinct contribution from the classification taxonomy in Tolmeijer et al. (2021).

Heyder et al. (2023) conducts a theoretical review on humans and AI agents interacting ethically in socio-technical systems for the sake of aligning values. They use this to produce a conceptual framework for the ethical management of Human-AI interactions. As in our review, they performed a qualitative analysis to identify patterns of themes that emerged in the reviewed articles. However, their review utilised a deductive approach with pre-existing codes and explicitly analysed the work through a lens of *sociomateriality*: the merging of technology and social phenomena in the emergence and observation of technology (Heyder et al. 2023). We instead use an inductive approach, and while our analysis incorporates many elements of agent-based approaches, our interpretation avoids using any specific theoretical framework.

Zoshak and Dew (2021) produced another survey paper on artificial moral agents, concluding that there was a dominating presence of Western ethical values in the design of these agents, at the expense of other theories. Similar to Heyder et al. (2023), they also used a deductive qualitative analysis approach, and similar to our paper, they used thematic synthesis to explore their results. Their work focuses on the normative aspects of value alignment, like Tolmeijer et al. (2021). This leads to a strong focus on ethical theories that is absent from our own work.

3 Methodology

We performed a structured literature survey and qualitatively coded the content of the included research articles. Qualitative coding analyses data by "reducing the data into meaningful segments and assigning names for the segments." (Creswell and Poth 2018 - 2018, p.183). In this case, the data was the content of the included papers. In this section we explain our search methodology. We discuss the results of our search, including the number of papers retrieved, in Section 3.1. A full description of our search terms, as well as the justification behind their use, is available in Appendix A.

The literature search was conducted in the Scopus database. Scopus covers leading AI publication venues including ACM, AAAI, and IEEE, and we supplement this coverage by adding papers from bibliographies. We limited ourselves to English-language papers due to a lack of translation capability. We selected our initial search terms to identify papers related to value alignment and human preferences. The relevance of additional topics

including virtue ethics, the social contract (Rousseau 2016), and multi-agent systems became apparent in the initial returned papers. In response, we performed an additional search with keywords on these new topics to find additional relevant papers. We filtered on publication year up to and including November 2023, the month when the search was executed. Our search thus creates a systematic coverage of the value alignment corpus up through November 2023.

We also filtered on subject area, only including papers tagged as computer science. We filtered on computer science to reduce the number of irrelevant papers from engineering or mathematics that, despite sharing similar themes with our exclusionary keywords, had not been removed from the search results. This was because of the broad scope of the term "value alignment", which was still matching these papers in engineering and mathematics. This did not exclude papers tagged with multiple disciplines that included computer science. We also filtered for final publication papers only. In total, our search returned 734 papers.

We selected papers to be included in the survey based on our inclusion/ exclusion criteria, which are fully specified in Appendix B. Our initial screening for papers to code had two phases: we examined the abstract and title in the first phase and then did a summary pass of the full text in the second phase. When inspecting the abstract and title, we excluded papers that used values in a non-socio-ethical sense, and we included those discussing incorporating human values or preferences into technology. When we investigated the full text of papers that passed the abstract/title examination, we included those that both attempted to define value alignment and discussed the challenges of getting humans and autonomous agents to behave in aligned ways, but excluded articles that only met one of these two criteria.

We performed an initial coding of the abstracts, introductions, and conclusions of the selected papers. This was done to generate initial codes quickly to frame our full coding in the later stage. During this process, we performed bibliography searches to add papers that were either relevant based on the cited information in the source paper or were highly cited across the papers under consideration. No further papers were added from the bibliographies of these additional papers in order to maintain a reasonable scope for this analysis.

From these initially coded papers, we applied a second set of inclusion/exclusion criteria that focused on including implementation-based papers over those focusing on the governance of AI or the moral status of autonomous agents. While these topics are relevant to value alignment, we left them for another analysis. We also excluded papers focusing only on a specific value, as value-specific issues were not the focus of this analysis. Finally, we included papers that discussed alignment in a normative sense in order to broaden our coverage.

We used *NVivo* for our coding. Coding was undertaken by a single coder. Coding consisted of assigning segments of the text or diagrams in the literature to codes that described the content of the segment in the context of characterising value alignment. We also coded for academic subjects that appeared in the text to see what subjects had been referenced in value alignment. Segments were generally passages of text rather than individual words or sentences, in order to retain the surrounding context in our coded samples. As per our inductive approach, we started with no codes and generated new codes as needed for categorising segments of text that felt relevant to the value alignment problem. All content in the papers we reviewed was considered for potential coding, except for titles, section headings and bibliographies.

Throughout the coding process, we grouped codes into categories and categories into themes, based on commonalities between their content. This allowed our understanding of value alignment to be synthesised from the common themes that appeared across the literature, supported by the work of different authors in the field.

3.1 Search Results

Fig 1 illustrates the literature search and screening process in a PRISMA flow diagram. Our initial search and screening resulted in 128 papers in which we coded the abstracts, titles and conclusions. 13 papers were omitted from consideration due to not being available in English. Searching the bibliographies from these papers led

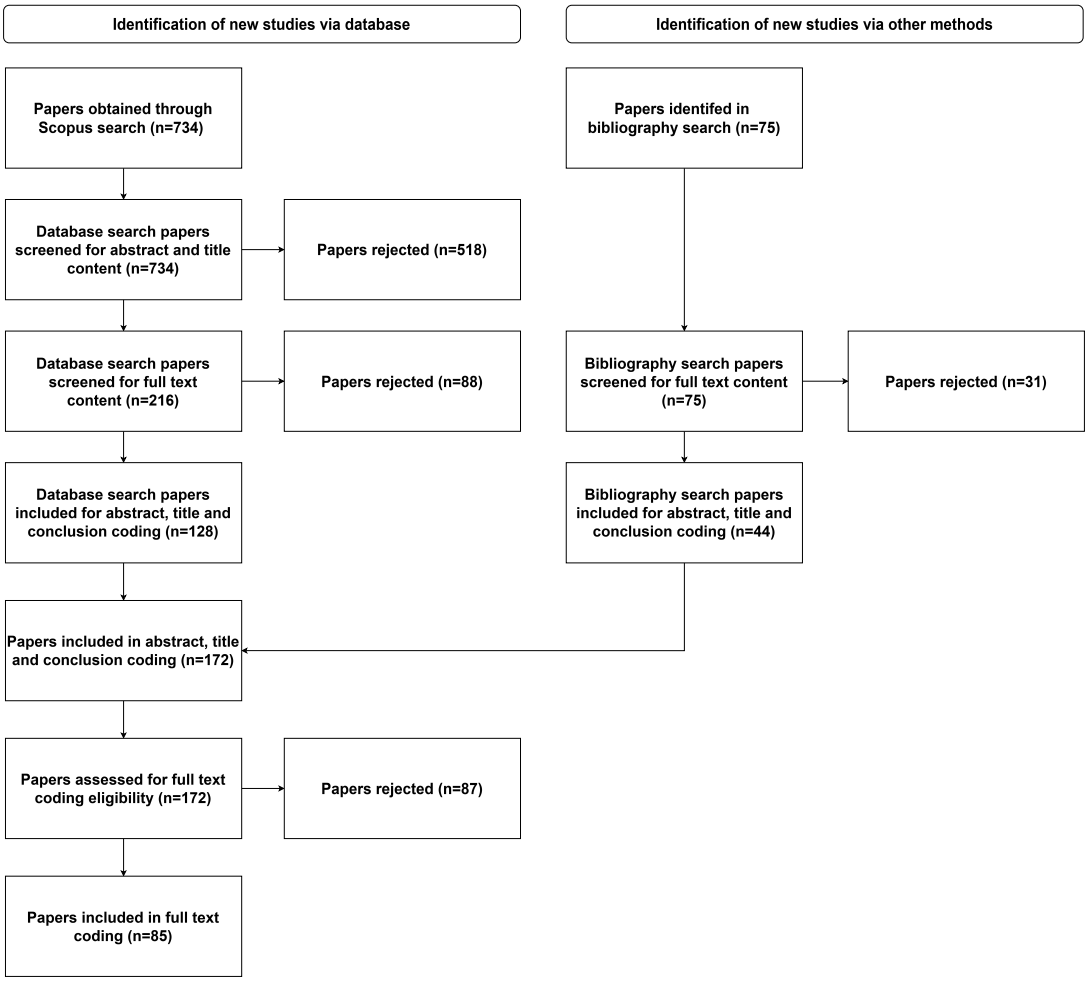

Fig. 1. PRISMA flow diagram (Page et al. 2021) illustrating our literature search results.

to the inspection of 75 additional papers, of which 44 were deemed relevant enough to include in the coding based on the full-text summary pass discussed in the previous paragraph. As stated, no further papers were added from the bibliographies of these additional papers. This resulted in 172 papers having their abstracts, introductions and conclusions used in coding. Of the 172 papers that underwent initial thematic coding, we selected 85 papers for full-text coding that focused on the implementation of value-alignment following our second set of inclusion/exclusion criteria (detailed in Appendix B.2).

While reviewing the literature, we categorised papers based on their content, using categories we created for this task. This was done so that we could observe the types of literature being published and how this changed over time.

- Extended abstracts were short proposals of future work.
- Research proposals were longer suggested several future agendas.
- Reviews examined previous literature.

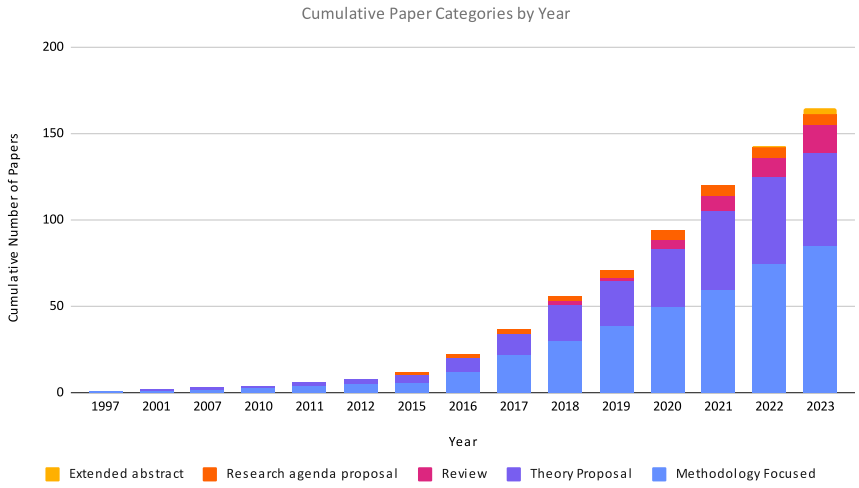

Fig. 2. Cumulative timeline of papers included in our survey, grouped by the category assigned to the paper. Categories are explained in Section. 3.

- Theory proposals were arguments for ways of thinking about the value alignment problem.
- Methodology-focused papers were based on the development of a specific value alignment technique.

The categorised papers in the review are charted in Fig. 2. We see that the field started publishing in earnest around 2015, with a generally increasing trend over time. To date, the field largely consists of developing methodologies and proposals for how to think about value alignment issues, with research proposals for next directions appearing on a mostly-annual basis. The split between theory and methodology varies by year, but overall, it would not seem to suggest any particular trends. In later years, we see reviews emerge as an increasingly popular category, suggesting increased attention in the field.

4 Analysis

Our coding of value alignment literature identified six themes that characterise value alignment in AI. We arrived at these themes through iterative refinement of the topic groupings, building our high-level themes from the bottom-up through hierarchical clustering of the individual codes. While there is overlap across these themes, we concluded from our iterative analysis that they formed a sufficiently concise and representative grouping of the themes demonstrated in the papers selected from the value alignment literature. It is interesting to note that the themes are organised along an axis from theoretical to applied content, though this was not a consideration during the clustering process.

These themes group the codes generated from the analysed literature in the following way:

T1. Value Alignment Drivers & Approaches focuses on understanding the motivations behind value alignment research, trends in approaches taken to studying the topic, and the range of subjects used in studying value alignment.

T2. Challenges in Value Alignment describes and discusses the identified challenges in accomplishing value alignment, in particular around how priorities are expressed and the challenges in reconciling different ethical systems.

T3. Values in Value Alignment highlights values themselves in value alignment research. It includes how values are embedded and represented in AI systems operating in diverse contexts, and how they change during a system's operation.

T4. Cognitive Processes in Humans and AI covers discussions on how humans use values in decision-making, including learning values and how they should be contextualised. It also discusses how these processes can be replicated in autonomous agents.

T5. Human-Agent Teaming discusses systems containing both humans and autonomous agents. It includes how these agents interact, and how information such as values and perceived system state is communicated in such systems.

T6. Designing and Developing Value-Aligned Systems groups information pertaining to the process of designing and implementing value-aligned systems in practice. This includes understanding the stakeholders of such systems and how the alignment of systems with stakeholders' needs can be tested.

Within these themes, we identified categories that describe the majority of each theme's content. We list these in Table 1. For each theme we also include the three papers that had the highest number of excerpts coded for the given theme. The fact that some papers, e.g. Stenseke (2023), appear as top papers in multiple themes demonstrates that some overlap exists between our themes.

To investigate the distribution of themes in the literature, after we constructed our themes we charted the number of papers from our survey that focused on each theme over time (Fig. 3). We counted a paper as focusing on a theme if at least 10% of samples coded from that paper were coded to the given theme. This is to exclude papers which may have one or two data samples coded to a given theme, but this amount is negligible compared to the more prominent themes discussed in the paper. Over time, the percentage of papers mentioning each theme has remained broadly consistent. This indicates that these themes are not just passing trends but are core open challenges for value alignment research. While the 10% threshold was chosen arbitrarily, testing other thresholds between 5 and 20% did not significantly impact the resulting distribution of themes.

We now analyse the themes and the literature they describe in greater detail. We have structured this discussion to best illustrate our understanding of value alignment developed through the analysis. This results in the first three themes (T1-T3) having their own dedicated subsections. For the themes of cognitive processes (T4) and human-agent teaming (T5), discussing them together as part of the value alignment process in systems proved more effective than discussing them separately. Given that cognitive processes describe components internal to the agent, while human-agent teaming looked at interactions between the two agents, there were naturally significant interactions between the two themes which were better analysed in tandem. Finally, our analysis led to many observations about the design and development of value-aligned systems (T6) across the themes, and we choose to highlight these where appropriate in other themes.

4.1 Value Alignment Drivers & Approaches

To characterise value alignment, it is essential to understand both its motivation and previous approaches to the topic. This reveals key objectives, concerns, and insights that have shaped the research.

4.1.1 Motivations. To better understand motivations for research on value alignment, we inductively coded our sample excerpts of given motivations found in the literature. The aim was to find common issues across these motivations. To illustrate our results we generated a word cloud of the 15 most prominent terms used in literature excerpts discussing motivations, grouping synonyms, shown in Fig. 4.

Combining this word cloud with an analysis of the text samples discussing motivation, we judged *autonomy* to be one of the most prominent drivers for researchers' concern with the value alignment problem. Vamplew et al. (2018) summarises the concern driven by autonomy as:

Table 1. Table of themes, representative categories and top three most relevant papers by number of coded data samples from that theme. The representative codes were taken from the highest-level groupings of codes within the given theme. The presence of papers in multiple themes demonstrates the interrelations between these themes.

Theme	Representative Code Categories	Top Papers by Samples-in-Theme
Value Alignment Drivers & Approaches	Motivations Technical and Normative Alignment Interdisciplinary Approach	1. Han et al. (2022) 2. Sutrop (2020) 3. Stenseke (2023)
Challenges in Value Alignment	Expressing Priorities Implementing Ethical Theory	1. Bench-Capon (2020) 2. Stenseke (2023) 3. Vamplew et al. (2018)
Values in Value Alignment	Context Value Dynamism Value Aggregation	1. Noriega, Verhagen, et al. (2022) 2. Han et al. (2022) 3. Liscio, Meer, et al. (2022)
Cognitive Processes in Humans and AI	Learning Reasoning & Decision-Making Self-Reflection in AI	1. Allen et al. (2005) 2. Cervantes, Rodríguez, et al. (2016) 3. Stenseke (2023)
Human-Agent Teaming	Sharing Knowledge Perception and Modelling Interactions Between Humans and AI	1. M. Cappuccio et al. (2021) 2. Poel (2020) 3. Sanneman and Shah (2023b)
Designing and Developing Value-Aligned Systems	Design & Development Process Stakeholders Value Alignment Testing	1. Noriega, Verhagen, et al. (2022) 2. Gabriel (2020) 3. Brown et al. (2021)

> Increases in the agent's intellectual capacity, the broadness of the actions available to it, and the breadth of the domain in which it is applied increase the difficulty in ensuring the agent's behaviour is aligned, and also the magnitude of the negative side-effects of any unaligned behaviour.

The autonomy driver can be broken down into several highlighted risks that were dominant in the literature. These included risks from *unpredictability*, *corrigibility*, and *negative impacts on human values* through the interactions of humans with autonomous agents.

Unpredictability as a risk stemmed from what Yudkowsky (2016) described as the problem of "unforeseen instantiation": humans' inability to properly anticipate all potential scenarios an autonomous agent may find itself in. As we endow autonomous agents with greater autonomy, the range of states they can explore, and the paths between these states, quickly become intractable to consider. Further complicating this is how an autonomous agent's view of the world is artificially constructed and may not appropriately capture our own understanding of the world. Lake et al. (2017) demonstrate this with the game *Frostbite*, where deep reinforcement learning agents learn through pattern recognition over game images, whilst humans rapidly build intuitive theories about goals, object types, and their interactions. This difference means that humans can flexibly adapt to new goals or

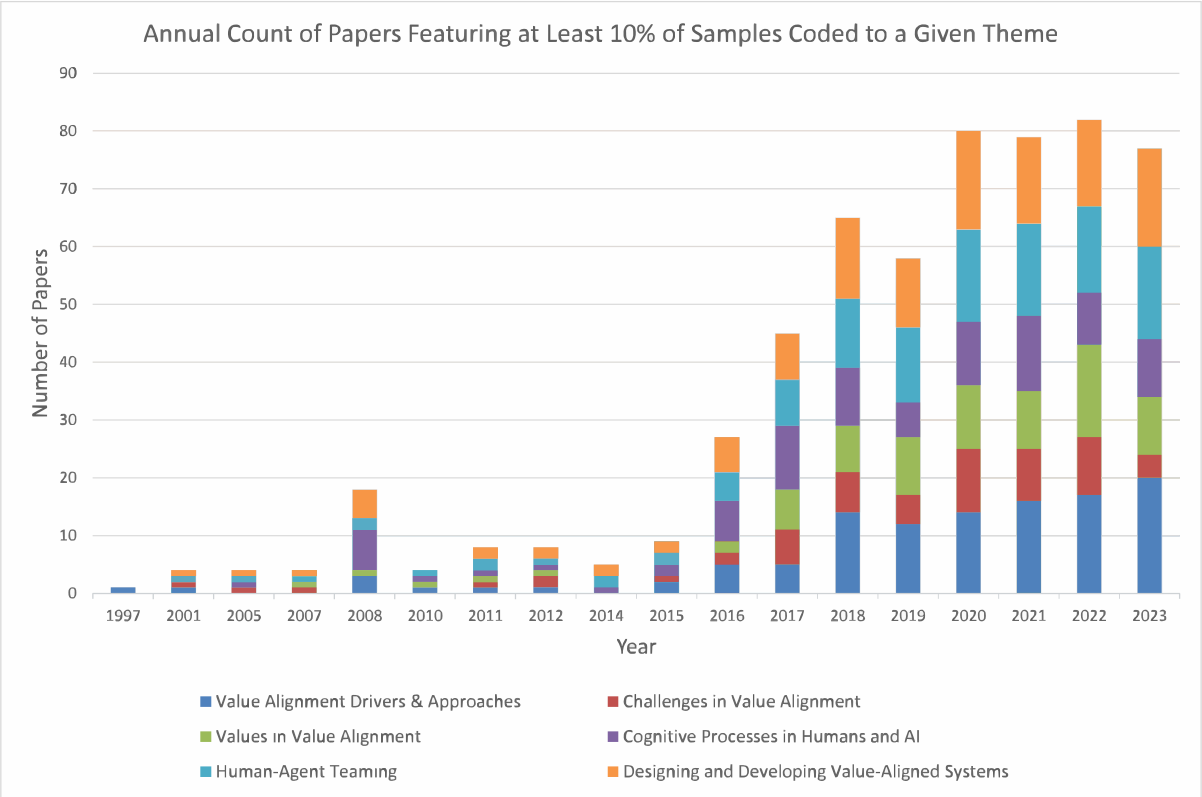

Fig. 3. Number of studies in our survey, grouped by theme, which contained at least 10% of its data samples coded to that group's theme, summed by year. We see that the representation of each theme remains broadly consistent, suggesting that these themes are characteristic of value alignment research.

Fig. 4. Word cloud of the 15 most frequent terms used in discussing the motivation to create value-aligned agents, grouping synonyms. We observed that *autonomous* as a grouped term had the highest weighted percentage across terms, ignoring terms that described system aspirations (e.g. moral; values; aligned), or descriptors of system components (e.g. systems; human; machines).

constraints, while the autonomous agent remains locked into its original objective and is unable to accommodate

even straightforward scenario modifications without complete retraining. This inflexibility contributes to the autonomy risk: due to the difficulty in fully envisioning all the different scenarios and their value-based risks necessary for training autonomous agents to act independently, agents may behave unpredictably when faced with situations outside their learned behaviour and potentially act misaligned as a consequence.

Prominent within discussion on the risk from unpredictability was concern regarding autonomous agents learning unintended strategies for obtaining rewards. This concept of *reward hacking*, where agents learn to exploit loopholes in their reward functions to achieve greater rewards, is regularly cited as a concern in the alignment literature (Sanneman and Shah 2023a; Zhuang and Hadfield-Menell 2020). This difficulty in anticipating how an agent may learn to solve a problem obscures the potential value impacts. Taken unintended strategy learning with unclear value impacts together, unpredictability not only complicates autonomous agent design and training, but it also obscures the scale of risk that misalignment presents.

The idea of incorrigibility, introduced by Soares et al. (2015), suggests that a sufficiently intelligent agent may not respond to attempts to correct its behaviour or shut it down, if it models doing so as counterproductive to maximising its utility function. If such a utility function were to be misaligned with the agent's stakeholders' values, then incorrigibility would pose a serious risk. This links to the concern around *superintelligence*, as conceived by Bostrom (2017). This was a frequently raised concern when discussing risk in value alignment, to the point where Bostrom was cited by approximately 25% of the papers included in the 172 papers in the first coding pass of our review. As Arnold et al. (2017) observed, the threat of uncontrollable AI agents strongly influenced the development of alignment initiatives.

Sarma et al. (2018) made the argument regarding alignment that there was no reason researchers could not pursue both near-term concerns and longer-term possibilities such as superintelligence, and that a continuum exists between these issues in the development of autonomous agents. The risk from misalignment comes from an agent's impacts on outcomes, rather than its capacity for cognition. So, while the notion of superintelligence has had a prominent influence on motivating value alignment and corrigibility concerns, it should not be seen as fully, or even largely representative of the problem.

Autonomous agents might also alter the values humans hold. Han et al. (2022) points out that it is nearly impossible to fully anticipate how technology will influence human actions, or the effect that this will have on a system of values held by humans. Svegliato et al. (2020) also acknowledges this concern regarding the unpredictable effects AI systems can have on stakeholder values, which they state stems from performing gradual modifications to an agent's reward function in the pursuit of achieving ethical behaviour. The idea of *moral deskilling*, that extended interaction with autonomous agents may harm our morals as humans, is the focus of Vallor (2015). Vallor concludes that the moral future of humans interacting with emerging technologies is not sufficiently understood, and greater attention needs to be paid to this risk. Given values changing over time and between contexts, which we discuss in Section 4.3.3, and the need to aggregate values, discussed in Section 4.3.4, it is evident that autonomous agents risk biasing human values in an undesirable direction, or reinforcing outdated values, should misalignment occur.

Where autonomy-based accounts emphasise agent capability and control, a second driver we identified in our analysis concerns the *embedding of AI within socio-technical systems*. Designing socio-technical systems with values in mind is not a new problem (Jasanoff 2016), nor is AI a new technology. However, recent developments have increased the scale, scope, and interdependence of AI-mediated decision-making. This shifts the nature of value alignment in socio-technical systems from involving bounded, task-specific systems with largely predictable and controllable effects, to dealing with more generalised and opaque AI agents whose behaviour may not reflect the designer's intentions even with explicit normative objectives specified. In such cases misalignment may emerge from learning dynamics or contextual shifts and hence is often difficult or impossible to anticipate.

Despite this, the presumed adoption of AI systems in decision-making capacities in society, and the delegation of higher-stake decisions to these autonomous agents, was rarely treated as a possibility rather than a certainty.

Sison and Redín (2023) described this as an 'inevitability' argument, though they argued against how truly 'inevitable' the need to use autonomous agents in moral roles was, as well as the 'inevitability' of humans choosing to develop autonomous moral agents. Behdadi and Munthe (2020) noted that the growing range of industries that autonomous agents were being deployed in, combined with their growing autonomy, increased the urgency of assessing the status of autonomous agents as moral agents. V. Dignum et al. (2018) made a similar observation but noted that the urgency was more in ensuring that autonomous agents acted in alignment with our ethics and values to build trust in the agents and ensure humanity's benefit, a sentiment echoed by Bench-Capon (2020) but with an emphasis on harm rather than trust.

A less mentioned risk in the given motivations, but one worth highlighting, was the political risk from AI. As Han et al. (2022) observes, the development of AI by a small group of actors risks a kind of "moral paralysis" when the values embedded in a sufficiently powerful AI are determined by a select few, causing significant impacts on those who don't share the embedded values. Soares (2018) points out that a sufficiently powerful system not only risks this moral paralysis but also concentrates enormous power to influence the future of humanity into a minority of people. The normative component of value alignment discussed in the next section, regarding which values should be built into the AI and who gets to decide, does little to help this concern.

Other common drivers for value alignment research included: a need for autonomous agents to act ethically (Badea and Artus 2022; Svegliato et al. 2020), and common values cited by policymakers such as safety (Chaturvedi et al. 2023; Juric et al. 2020), trust (V. Dignum et al. 2018; Firt 2023), and fairness (Osoba et al. 2020). Hagendorff (2020) also provides a further list of risks from autonomous agents that they considered neglected by modern guidelines for developing such systems.

Taking these motivations together, we can summarise that the objective of value alignment research is to manage the risks associated with autonomous agents (unpredictability, corrigibility, and impacts on human values from human-AI interactions) in such a way that the autonomous agents can be allowed to impact our values in a positive manner, while also accommodating the inherent political and ethical risks associated with value influence.

4.1.2 Technical and Normative Alignment. Value alignment is often described in terms of both normative alignment problems and technical alignment problems (Chaturvedi et al. 2023; Gabriel 2020; Sutrop 2020). The former refers to the challenge of deciding what values and principles a system should be aligned with, while the latter refers to the challenge of getting the system to align with these values. Firt (2023) also suggested a third dimension, calibration problems, which focuses on fine-tuning systems for specific values for trust building. However, Firt presupposed in their discussion that this would become relevant after the first two problems were solved.

While this separation between normative and technical alignment is convenient for narrowing the scope of a given research paper, it does obscure the interaction between these two sides of the same coin. As Tolmeijer et al. (2021) states, "In the context of implementing machine ethics, it can be a pitfall for philosophers to use a purely theoretical approach without consulting computer scientists, as this can result in theories that are too abstract to be implemented. Conversely, computer scientists may implement a faulty interpretation of an ethical theory if they do not consult a philosopher.". Additionally, Peschl et al. (2022) observed that the value alignment field has largely focused on reward learning, a technical problem, as demonstrated by the prominence of Hadfield-Menell, Dragan, et al. (2016) throughout our analysed papers. Research exploring the normative side was comparatively rarer in our survey, mainly discussed at higher conceptual levels (Butlin 2021; Gabriel 2020) rather than developed into actionable processes.

However, Robinson (2023) describes this normative aspect as crucial to good alignment: "The first step is to find principles that the AI agent possibly could and should be aligned with. There is no point trying to align AI agents to principles they cannot be aligned with, or that we have no interest in aligning them with.". As Sutrop (2020) cautions, AI developers risk assuming that a normative solution will naturally follow from sufficiently technically

capable AI, even though we as a society still remain undecided about our value priorities. This also links into the need to aggregate our values across more groups and cultures than just the developers of such systems to enable alignment, which we discuss further in Section 4.3.4. Focusing on the technical side of alignment, as was indicated in our analysis, at the expense of the normative elements risks neglecting important aspects of the process.

4.1.3 Interdisciplinary Approaches. An important observation from examining previous value alignment work is that value alignment is often seen as an *interdisciplinary* challenge, a sentiment echoed by authors including Dafoe et al. (2021), Khan et al. (2022), Noriega, Verhagen, et al. (2022), Russell et al. (2015), and Sutrop (2020), and Li (2021). Numerous subjects contributed to the value alignment research literature, which we demonstrate in Table 2 by citing a range of contributions from different disciplines. Many of these subjects were themselves interdisciplinary, further expanding the list of subjects relevant to the value alignment problem. This demonstrates both the importance and challenge of integrating a diverse body of knowledge into value alignment research and the development process of value-aligned systems.

The contributions of the subjects most frequently mentioned as being relevant to value alignment are well summarised by Dafoe et al. (2021): "psychology, to understand human cognition; law and policy, to understand institutions; history, sociology and anthropology, to understand culture; and political science and economics, to understand problems of information, commitment and social choice.". In further detail, we see psychology, neuropsychology and cognitive sciences suggested for understanding values and value structures in concept and in action (Butlin 2021; Cervantes, Rodríguez, et al. 2016; Fisac et al. 2020; Haas 2020; Montes et al. 2023; Ratoff 2021; Sarma et al. 2018); philosophy and legal studies for understanding values, moral hazards and explainability (Hadfield-Menell and Hadfield 2019; Zoshak and Dew 2021); and humanities subjects applied to understanding human decision-making, cooperation and value aggregation (Holgado-Sánchez et al. 2023; Lera-Leri et al. 2022).

The main barrier to interdisciplinary cooperation discussed was translating ideas between domains. For example, Cervantes, Rodríguez, et al. (2016) noted that emulating the brain's complex decision-making process in autonomous agents was enormously difficult, in part due to the lack of an accurate formalism of human brain functions that could be replicated in autonomous agents. This limits our ability to exploit neuropsychology results in AI. Rahwan (2018) discusses the cultural divide that exists between engineering and the humanities: while practitioners of the humanities are skilled at considering the applications of moral hazards, they can struggle to articulate them in ways that engineers can operationalise. Similarly, engineers are not always able to quantify their own work in a way that can be understood by humanities practitioners. Given the need to integrate normative and technical aspects simultaneously, this cultural divide presents a significant barrier.

4.2 Challenges in Value Alignment

We have examined the difficult nature of doing value alignment in Section 4.1.1, given the opaque nature of autonomy and in the potential scale of value impacts resulting from the inevitable embedding of autonomous agents. In this section, we expand on the particular challenges highlighted by our analysis, from the perspective of an automated process involving interacting humans and autonomous agents.

4.2.1 Expressing Priorities. Throughout the literature, there were repeated efforts to make values known by both autonomous agents and humans. In the former case, this was predominantly in the form of these agents learning from humans, while the latter involved enabling humans to both inspect values and understand their own desires.

During our analysis, we observed that values, goals and preferences were used interchangeably as targets for alignment throughout the literature. This observation was also made by Gabriel (2020) and Zurek and Mokkas (2021), with the former pointing out that these concepts are not equivalent. Our analysis would suggest that the terms can be interpreted as follows:

Table 2. Interdisciplinary perspectives in value alignment literature.

Discipline	Papers
Cognitive informatics	Cervantes, Rodríguez, et al. (2016)
Cognitive Science	Ratoff (2021) Fisac et al. (2020)
Economics	Hadfield-Menell and Hadfield (2019)
Game Theory	Davoust and Rovatsos (2020) Gavidia-Calderon et al. (2022)
Legal theory	Hadfield-Menell and Hadfield (2019)
Neuroscience	Crook and Corneli (2021)
Neuropsychology	Sarma et al. (2018)
Philosophy	Han et al. (2022) Maitra (2020) Allen et al. (2005) Stenseke (2023) Zoshak and Dew (2021)
Psychology	Butlin (2021) Mechergui and Sreedharan (2023) Govindarajulu et al. (2019) Montes et al. (2023)
Social choice theory	Lera-Leri et al. (2022)
Social psychology	Waser and Kelley (2018)
Systems engineering	Aliman, Kester, et al. (2019)

- Values, existing at the most general and abstract level, are used in forming desirable goals in more specific contexts and evaluating possibilities.
- Goals are formed from the values manifesting in the given context, with goals constructed in pursuit of those values. The values that have led to a goal being chosen will not necessarily be obvious when auditing the goals, however.
- Preferences can be understood in the conventional sense: as priorities between different options and outcomes based on our underlying goals and values, as well as contextual priorities between goals and values themselves.

Because of the connection between values and goals (Schwartz 1992), the challenges in expressing our goals are inherently tied to the value alignment problem. Hence, the difficulty in goal expression is an essential issue in the value identification problem. Hadfield-Menell and Hadfield (2019) in particular describes the problems in expressing goals as being entirely responsible for the alignment problem. While we would not want to point to goal misspecification as the sole root of challenges in value alignment, as we find this too reductionist, it is certainly worth continued attention.

Given that many of the goals we would ideally use AI agents for are complex or hard to define (Christiano et al. 2017), it seems inevitable that we struggle to define these goals correctly or fully. For example, Thornton et al. (2017) points out that, ideally, we want self-driving vehicles to drive safely and smoothly, but this requirement does not translate into machine language easily. Hadfield-Menell and Hadfield (2019) also agrees that the misspecification of AI reward functions is often unavoidable. Many of the challenges in expressing goals can be described in terms of under-specifying goals (Alamdari et al. 2022), with our stated goals failing to capture important nuances.

This underspecification often occurs in the form of developers attempting to consolidate goals into utility functions, a formula for capturing the preference of one outcome over another. However, these utility functions often fail to capture what we really desire (Aliman and Kester 2019). As Vamplew et al. (2018) notes, maximising expected utility functions, the predominant paradigm in modern machine learning architectures, often leads to unexpected behaviour that fails to align with the designer's original goal. While a popular approach, innovation is still needed to consider alternative mechanisms beyond utility functions to encode our values.

Expressing values and preferences directly, circumventing goals in the process, is also a challenge. Humans lack adeptness at expressing their preferences in explicit statements (Bharadhwaj 2021; Liscio, Lera-Leri, et al. 2023; Milli et al. 2017) or quantified forms (Rosenthal and Veloso 2012), and struggle to mentally compare sufficient scenarios to cover all eventualities (Rosenthal and Veloso 2012; Visser et al. 2011). Preferences also vary between individuals, complicating preference modelling by autonomous agents due to the need to be able to adapt to differences in preferences, or to aggregate them (Rosenthal and Veloso 2012). People face similar challenges in expressing (Sanneman and Shah 2023b) or evaluating (Liscio, Meer, et al. 2022; Sanneman and Shah 2023a) values, struggling to express values outside of concrete examples (Liscio, Lera-Leri, et al. 2023) as a result of their abstract, contextual, and often incommensurable nature.

Taken together, the difficulties in expressing values, goals and preferences illustrate one of the fundamental challenges in value alignment: expressing our values in the first place, either as values, goals, or preferences, is hard. Research explicitly attempting to improve how values, goals and preferences can be expressed was not found in the survey, presenting a possible opportunity for the future. If we could define our values, goals or preferences more accurately, this would help to reduce the noise in communicating these concepts to autonomous agents while generating evidence of what has been shared in the learning process, making alignment a smoother and more scrutable process.

4.2.2 Implementing Ethical Theories. The other most prominent challenge in the value alignment literature reviewed was the difficulty of summarising and translating the millennia of ethical thought into a machine-compatible format. Ethical systems serve as vital repositories of our developed values as groups of humans, and no good argument for not making use of them in aligning humans and autonomous agents emerged from the analysis. However, different systems of ethics present different frameworks for modelling values in autonomous agents and have implications for the alignment process.

Within the reviewed papers, the ethical discussion centred around three Western theories: consequentialism, usually in the form of utilitarianism; deontology; and virtue ethics. M. Cappuccio et al. (2021) provides a succinct summary of the three theories:

> Consequentialism proposes to evaluate an action's moral value based on the utility created by the action's effects. Deontology, in turn, focuses on the intentions that motivated the action and judges

> whether or not the agent acted out of authentic goodwill to fulfil their obligations and respect others' rights. . . . Virtue Ethics differs from both Consequentialism and Deontology because, unlike them, it does not primarily ask whether an action produces desirable effects or is motivated by good intentions, but whether a person deserves praise or blame and whether the kind of life they live is worth living.

Consequentialism and the related utilitarianism have thrived under machine learning paradigms. Credit is given to the popularity of utility functions and reinforcement learning (Gabriel 2020) and its pre-existing history with economics (Vamplew et al. 2018). Franzke (2022) also claims that utilitarian values pervade AI governance. As a result, deontology-based approaches have been left feeling marginalised, with most of the papers identified as deontic in our survey being theoretical analyses rather than implementations (Pagallo 2016; Rahwan 2018), though some deontic ideals seem to manifest in reasoning-based implementations (Cranefield et al. 2017; Szabo, Such, and Criado 2020). Meanwhile, virtue ethics has been championed as an under-represented (Murray 2017; Vamplew et al. 2018) but robust foundation for value alignment (M. Cappuccio et al. 2021; Franzke 2022; Pagallo 2016), but concrete implementations of the framework remain in their infancy (Crook and Corneli 2021; Govindarajulu et al. 2019; Stenseke 2023).

Even if priorities and ethical theories are able to be successfully expressed and implemented, there remains the design challenge of choosing between them. As we mentioned in Section 4.1.2, technical value alignment work often ignores the normative aspect. However, in choosing (implicitly or explicitly) an ethical theory and the goals and values for a system to align to, a value-laden decision is itself being made about which values and goals are worth empowering through artificial intelligence systems (Davoust and Rovatsos 2020). As stated by Soares (2018), "human goals are complex, culturally laden, and context-dependant", demonstrating how our choice of goals incorporates our own cultural biases. Given the political risks of artificial intelligence discussed in Section 4.1.1, these decisions must be made carefully if we want to avoid marginalising schools of ethical thought or particular values, particularly given the dominance some cultures currently enjoy in AI development.

Comparisons between ethical theories pervade the literature. To quote Bench-Capon (2020): "it is seen that each approach has its own particular strengths and weaknesses when considered as the basis for implementing ethical agents, and that the different approaches are appropriate to different kinds of system.". We list the strengths and weaknesses that we identified for different frameworks in the surveyed literature in Table 3. This table highlights that the value alignment literature does not converge on a single ethical theory as being optimal for value alignment but rather reflects context-dependent and sometimes conflicting claims about the suitability of different ethical theories as frameworks for modelling values in artificial agents.

Given the individual strengths and weaknesses of different approaches, an instinctual response might be to assume that no single theory can suit all value alignment needs. Such an observation has been made by several authors (M. Cappuccio et al. 2021; Pagallo 2016; Vamplew et al. 2018). As Gabriel (2020) puts it: "it is very unlikely that any single moral theory we can now point to captures the entire truth about morality. Indeed, each of the major candidates, at least within Western philosophical traditions, has strongly counterintuitive moral implications in some known situations, or else is significantly underdetermined.". Given the role of ethical systems in representing groups of values, we can infer from this statement that no single ethical system used to perform value alignment would suit all possible situations.

Considering the inadequacy of a single moral theory, the notion of combining ethical frameworks has gained popularity within the value alignment literature. Drawing from Allen et al. (2005)'s initial rationale for hybrid ethical learning systems, researchers acknowledge the merits of combining ethical frameworks to alleviate the weaknesses of individual ethical theories. Some practical examples of hybrid ethical learning systems can be seen in Córdova and Vicari (2022) and Thornton et al. (2017) and Cervantes, Rodríguez, et al. (2016). In these papers different ethical frameworks, particularly deontology and consequentialism, are combined to model complex ethical behaviour in example applied scenarios. However, Stenseke (2023) cautions about cherry-picking in this

Table 3. Commonly reported strengths and limitations of major ethical theories as discussed in the value alignment literature. The table summarises claims made by different authors, which are not always mutually consistent and often depend on assumptions about scale, context, and implementation.

Ethical Theory	Claimed Strengths	Claimed Limitations
Consequentialism & Utilitarianism	Often described as formally specifiable and mathematically grounded (Stenseke 2023); can provide precise action guidance under well-defined objectives (Murray 2017); lends itself to optimisation-based implementations and transparency of decision criteria (Bench-Capon 2020).	Criticised for difficulty handling incommensurable or uncertain values (Allen et al. 2005; Eckersley 2019); vulnerability to reward hacking and specification gaming (Christiano et al. 2017; Sanneman and Shah 2023a); high computational demands in realistic settings (Allen et al. 2005; Bench-Capon and Modgil 2017); challenges in assigning comprehensive and stable utility functions (Mechergui and Sreedharan 2023; Thornton et al. 2017; Yudkowsky 2016); limited descriptive fit with human moral reasoning (Atkinson and Bench-Capon 2016; Gavidia-Calderon et al. 2022).
Deontology	Valued for rule-based clarity, predictability, and ease of enforcement (Bench-Capon 2020; Stenseke 2023); supports coordination and cooperation by constraining permissible actions (Pagallo 2016); can offer clear prohibitions and obligations (Murray 2017).	Relies heavily on the quality and completeness of rules (Bench-Capon 2020; Murray 2017; Pagallo 2016); prone to normative conflicts between rules (Bauer 2020; Bench-Capon 2020); limited flexibility in novel or unforeseen contexts (Bauer 2020; Constantinescu, Voinea, et al. 2021); requires substantial background knowledge for interpretation (Allen et al. 2005; Badea and Artus 2022).
Virtue Ethics	Emphasises contextual and socially embedded ethical behaviour (Franzke 2022; Pagallo 2016); supports moral development, learning, and adaptation over time (M. L. Cappuccio et al. 2020; Coleman 2001; Reichberg and Syse 2021; Stenseke 2023); often associated with explainability and hybrid human–machine ethical systems (Bench-Capon 2020; Gamez et al. 2020).	Widely regarded as difficult to formalise or implement (Govindarajulu et al. 2019; Stenseke 2023; Sutrop 2020); lacks consensus on relevant virtues (M. Cappuccio et al. 2021; Pagallo 2016); provides limited action-level guidance (Murray 2017).

approach, where suitable features are chosen from ethical theories without considering them in the context of holistic ethical cognition, as it risks neglecting to consider how these features relate to the wider reasoning of agents in complex dynamic contexts. While the three previous examples showed promising results within their domains, they all lack empirical validation in a real-world holistic context.

On a final note, Svegliato et al. (2020) suggests that a chosen ethical framework may be incompatible with the goals of the system if a solution does not exist within the chosen framework. Rather than being seen as a failure of

the system, it should instead draw the designer's attention to the ethics of the task itself, and whether it is worth pursuing. Davoust and Rovatsos (2020) calls this a social dilemma: a conflict between what is good for the wider group and what the autonomous agent (or its owner) wishes to achieve. In this sense, ethical frameworks, while not always simple to work with, can serve as important ethical sense checks for developing value-aligned systems. To develop ethical, value-aligned agents, the choice of an ethical framework should be guided by aspirations rather than practicalities. To ensure accountability and transparency in this process, the design decisions should be thoroughly documented, including a clear justification for the selected ethical system(s).

Summarising, ethical schools of thought are a highly relevant topic in value alignment, but their different aspects present challenges for both selection and implementation. At the same time, relying too much on a single framework risks leaving its particular weaknesses unresolved and the advantages of another framework unused. This challenge of exploiting and combining ethical frameworks, already used to structure our daily society's functioning, is a key obstacle in making use of our existing value-based knowledge in the creation of autonomous agents that can integrate with said society.

4.3 Values in Value Alignment

We now reach the central concept of value alignment as a topic. The abstract nature of values necessitates processing them before they can be used by digital agents. For example, translating goals and preferences discussed in Section 4.2.1 is one form of this processing. Here, we examine the nature of this value processing.

It is important to establish that most reviewed papers based their interpretations of values on the work of Schwartz (2012, 1992). While some reference was made to other values scholars like Rokeach (1967) and his work on value classification, and the work by Graham et al. (2013) on moral foundations theory, these were in the minority. It is easy to understand why this is the case: Schwartz's theory is well supported by empirical evidence, and its structure lends itself easily to machine implementation. Regardless, the applicability of other interpretations of value systems in value alignment remains under-explored. It is important to consider when developing specific value-aligned systems whether Schwartz's model is a suitable choice for the system's purpose, or if an alternate theory would be more appropriate. In particular, psychology and ethical theories interpret values in different ways, and this distinction may be critical to consider depending on the goal of the work.

As mentioned in the introduction, values can be understood as latent drivers in evaluating situations and outcomes. Possessing a combination of contextualised values can lead to positive, negative or indifferent assessments of a presented situation. This is a simplified characterisation, but it will support the remainder of this section.

4.3.1 Stakeholders. So far we have discussed value alignment in the context of autonomous agents aligning with human values, making human stakeholders the primary sources of values that a system needs to align with. Values can originate in individuals, but in the literature they are more often discussed as resulting from the aggregation of particular groups, such as developers or users of AI systems. While stakeholder groups have been identified (see Tomsett et al. (2018) and Meske et al. (2022)), we found developers and users to be the main focus.

Our analysis reveals that effective alignment requires a two-pronged approach: first, *understanding a stakeholder's values* so that they can inform an agent's actions, and second, *empowering those values* to ensure the agent adds value and avoids undermining existing ones. This relates back to both the relevance of an agent's influence on values from Section 4.1.1 through value empowerment, and the difficulties in expressing values from Section 4.2.1 when trying to understand a stakeholders' values.

Understanding stakeholders was often framed in terms of understanding stakeholder values and the variance between stakeholders. The technical approach to this was primarily reinforcement learning and utility functions (Ficici et al. 2008; Hadfield-Menell, Dragan, et al. 2016; Zintgraf et al. 2018), but other methods such as CP-nets (Loreggia et al. 2019), or multi-label classification with orderings over values (Siebert et al. 2022) were also

observed. Engineering approaches were also used, including requirement engineering from software engineering (Liscio, Meer, et al. 2022) and guided annotation (Liscio 2021).

Liscio, Lera-Leri, et al. (2023) states that this stakeholder value inference process cannot be performed entirely through computational means, however, since behaviour alone may not reveal enough about values. Gabriel (2020) also raises the question of autonomous agents following instructions, versus these agents following implicit stakeholder needs, which brings back the issues in expressing goals for value alignment from Section 4.2.1. Siebert et al. (2022) showed improved performance at estimating value rankings of users when combining stated motivations with choices made as information for the autonomous agent, showing the merits of combining explicit and implicit information. Still, the gain was relatively small compared to the effect of only inferring over given motivations. This might also suggest a difference between the values driving user choices and the values users report when asked to justify their choices, or at least an incompleteness of the latter, further complicating this inference process.

On the topic of empowerment, Feffer et al. (2023) raises the imperative that the AI ethics community has to empower stakeholders of autonomous agent systems in the co-creation of such systems. The authors, along with Siebert et al. (2022), call for a more participatory design of these systems. Vallor (2015) argues for a further form of this, desiring a cultural shift to bring all members of society as collectively responsible in the design process, rather than leaving it in the space of product designers and marketers. Aler Tubella, Theodorou, V. Dignum, et al. (2019) supports aspects of this notion, suggesting the value of educating the population about their ability to shape the development of society through responsible AI technology. Liscio, Lera-Leri, et al. (2023) also points out the importance of ensuring stakeholder consent and transparency about the fairness of the process. While they discuss this in the context of value inference, it relates to the idea of stakeholder empowerment through making them active decision-makers in the process of interacting with these systems, rather than just sources of data.

These papers indicate that empowerment is not just about designing autonomous agents to satisfy values, but also about incorporating more stakeholders into the design process so that these values can be properly recognised in the first place. It should be understood that stakeholders in value alignment are not just sources of information about values, but active participants in the system aiming to achieve a value-aligned state.

It is also worth calling attention to existing research in designing autonomous systems with stakeholders and values in mind. Value-sensitive design (Friedman 1996) and participatory design (Muller and Kuhn 1993) are active areas of research that offer concrete methodologies and case studies grounded in real-world deployment. They also treat stakeholder identification, value elicitation and empowerment as core design activities, rather than downstream problems. Similarly, applied domains such as recommender systems have long examined multi-stakeholder value trade-offs without framing them explicitly as value alignment problems (Smith et al. 2023). Engaging with these practical approaches can help ground alignment research in real-world constraints, complement purely theoretical or moralising discussions, and reduce the risk of reinventing the wheel.

4.3.2 Contextualisation. The key process in transforming values from "abstract motivations that guide our opinions and actions" (Schwartz 2012) to decision-making factors is *contextualisation*. Contextualisation is the interpretation of values in a given context and representing them through contextual proxies so that they can be acted upon by autonomous agents.

Values are contextualised by several means in the literature, including:

- Goals (Badea and Artus 2022; Cranefield et al. 2017; Montes et al. 2023) consistent with the definition from Schwartz and Bilsky (1987) that values are grounded as goals.
- Norms (Aler Tubella, Theodorou, F. Dignum, et al. 2019; Pigmans et al. 2017; Zurek, Araszkiewicz, et al. 2022) as representations of grouped values.

- Design features in the agent's environment that could be interacted with directly, or measures of some combination of these interactive features (Cranefield et al. 2017; Noriega and Plaza 2022), in line with the work by Van de Poel (2013) on embedding values through value hierarchies.

Value hierarchies in particular illustrate how contextualisation relates to the *embedding* of values in autonomous agents through design. Noriega, Verhagen, et al. (2023) outlines a generalised heuristic approach to contextualisation:

> [Online Institution (OI)] values can be contextualised and embedded in four successive stages: (i) values for the application domain and [Conscientious Design] categories for the consensual preferences of the three design stakeholders towards the OI, (ii) for the individual preferences of each of the design stakeholders of the OI; (ii) then for the compatibility requirements of the situated OI; and, finally, (iii) for the six [World; Institution; Technology]-articulation design concerns (abstraction, grounding, specification, implementation, input and output).

In this heuristic we see contextualisation, and in tandem context, described in terms of values, applications, stakeholders and institutions, and design constraints. As we might expect, a general context structure is very difficult to define, as understanding the influence of various factors on decision-making is a problem that has challenged decision scientists for decades (Spektor et al. 2021). That said, we concur with Noriega, Verhagen, et al. (2023) that the concepts given in the quote (abstraction, grounding, specification, implementation, input and output) are effective starting points to consider in approaching value contextualisation in practice.

This process of contextualisation is also discussed by Poel (2020), though not in explicit words. In this paper, the author emphasises the nature of embedded values in an autonomous agent and how the agent should adapt to different contexts to better realise these values. Here, we see values treated as something internal to the agent: intended objectives embedded through design that need to be approached differently depending on the context. This contrasts the idea of treating values as concepts held externally to an agent by stakeholders, which are interpreted by the agent to produce different objectives in relation to contextual proxies in the environment. This would effectively make the agent's only value satisfying its stakeholders, a flexible but unpredictable objective. Considering the risks associated with autonomy discussed in Section 4.1.1, a safer design approach would be to embed clear, well-defined values and contextualization methods for the agent and its associated task. This would promote predictability and auditability.

We see from these papers that contextualisation is a process impacting both the design and ongoing operation of an autonomous agent. The sheer diversity of contexts and the fact that they are difficult to fully anticipate in advance make programming these in advance impractical, if not impossible, as we alluded to in Section 4.1.1 when discussing autonomy. A means of expressing contexts accurately to autonomous agents in an ongoing fashion, an extension to the challenge of expressing values, and methods for autonomous agents to learn contexts on their own, would be essential for reliable value-aligned systems.

4.3.3 Value Dynamism. Values are formed by human stakeholders and then contextualised by the agent's operating conditions to impact the agent's decision-making. This leads to *value dynamism*: the changes in values over time and the timescales these changes occur at. This dynamism implies a need for value alignment to be able to react to these changes, rather than only aligning with an initial state of values.

As discussed in Section 4.3.2, contextualisation is a core process in how values are interpreted. Because contexts have varying levels of granularity and are defined by complex circumstances, context changes can rapidly occur during an agent's operation and, hence, change how values are interpreted. These types of value changes occur over the shortest time frames: context change can occur in as short a time as between individual decisions. For instance, what values impact a social care robot's interactions with a patient may need to be very different to those informing interactions with a care professional or family member. Said robot could need to switch behaviours in

the time it takes to move between the rooms these individuals are in. The frequent and unpredictable nature of context changes, combined with the very granular and diverse nature of contexts, poses one of the most challenging aspects of reliable value alignment.

Stakeholder composition reflects the goals of a system's stakeholders and the means of aggregating stakeholder values. Stakeholder goals and values change over time, which may result in misalignment with autonomous agents, even if they were aligned previously (Chaturvedi et al. 2023). For example, a user of social media may originally prioritise their entertainment, which the social media platform's recommender engine is aligned with by providing content the user finds entertaining. Later, the same user wishes to be more mindful about their use of social media and instead prioritise their mental health by not clicking through content carelessly, even if it would be entertaining. The recommender system must adjust to the user's new goals and priorities to remain aligned, rather than continuing to provide endless entertainment.

To handle this form of dynamism, a regular assessment of the system's understanding of its stakeholders is necessary. This again highlights the need raised in Section 4.3.1 to engage with stakeholders as active participants in the system but raises additional questions. How regularly this assessment needs to occur is not clear, and concerns about the user's privacy and cognitive demands on them are highly relevant. Some insights about modelling stakeholder composition change could potentially be gleamed from the literature on norm emergence (Gelfand et al. 2024; Morris-Martin et al. 2019), given the role of norms as representing grouped stakeholder values.

In the long term, the contextualisation of values changes over time because society evolves. Han et al. (2022) cites the example of men duelling in centuries prior, as this was the appropriate response to their honour being threatened, but such practice would be illegal in many countries these days. While we still understand the concept of honour as a value, we practice it differently in modern times. It would also be inappropriate to assume that this societal change will not continue in the future, as new generations develop new thoughts on morality and norms. Hence, we should acknowledge that systems may need to adapt to changing values in a greater sense than reconsidering the stakeholders of concern or being contextually adaptive.

That said, the length of time over which *societal value changes* occurs seems to be quite long, in the span of years, decades, or even centuries. The nature of systems, autonomous agents, and our concept of value alignment will no doubt further evolve in that time. Good practice for this kind of misalignment is to enable the evolution, or if necessary replacement, of systems using autonomous agents when needed, and avoiding these systems becoming too ingrained into any particular process, such that changing or removing the system proves problematic. While this cultural shifting of values may not seem like a significant risk factor, the abundance of legacy systems employed in modern society cautions us not to take the ability to swap out a system for granted, particularly as that system's capabilities become more impactful.

4.3.4 Value Aggregation. When inevitably dealing with multiple stakeholders, *value aggregation*, the reconciliation of diverse stakeholder interests to determine system objectives and considerations, becomes necessary. This aggregation needs to occur both on a large scale, agreeing cultural values and the means to implement them in autonomous systems, and on a smaller scale, aggregating values between individuals or smaller groups where cooperation is necessary for alignment.

While examining the literature, the majority of papers only looked at what Liscio, Lera-Leri, et al. (2023) refers to as *micro value alignment*, which is alignment occurring between two agents. Liscio et al. defines *macro value alignment* as leading to alignment between more than two agents through the emergence of mechanisms like norms. Particularly egregious was a lack of attempts to align multiple utility functions, given their prominence as a value encoding mechanism.

Since plenty of research exists on norm construction in multi-agent systems, a macro alignment mechanism according to Liscio, Lera-Leri, et al. (2023), then it is also possible that this work on norm construction is not

being properly integrated into the value alignment literature. This could be because this norm construction work is not referred to as value alignment, or because the methods are not based on neural systems, and hence not considered by researchers defaulting to a machine learning approach. Given the extensive amount of research in this norm construction space, integrating this work into promising value alignment approaches would be an effective research direction, in line with what Allen et al. (2005) described as hybrid systems.

Complicating aggregation is the fact that values vary enormously between stakeholders, and often in incompatible ways (Butlin 2021). This produces what Haas (2020) refers to as a moving target for aggregation. Combining the diverse interests of stakeholders becomes more challenging as the range of relevant values increases and fully satisfying everyone is often impossible in many situations. This is particularly true in the presence of conflicting values and their associated goals, which is often the case. For example, system providers may be able to create a more optimised system with more data and, hence, generate more profit, but users want their privacy to be respected. Social choice scholars such as Arrow (2012) and Sen (n.d.) have written at length about the various impossibilities in achieving a complete satisfaction of all values between stakeholders, particularly if a utility-based approach is adopted (Aliman and Kester 2019).

These unavoidable conflicts make value aggregation itself another value-laden process, much like selecting an ethical system. Trade-offs are a natural part of decision-making, particularly in the Schwartz model where some value categories, such as security and self-direction, are inherently opposed. However, trade-offs must still be decided regarding whose and which values are prioritised in a given context. When aggregating these different values for the sake of decision-making, it is plausible that different value aggregation methods will benefit different stakeholders and values in different ways. It may also not be possible to optimise a system towards one set of values without neglecting another. This could pose serious risks if the trade-offs made are not carefully monitored and made explicit, such as trading one group's safety or privacy for another group's profit or power.

In the literature, autonomous agents are often treated as having the single objective of aligning with a human. In practice, both autonomous agents and humans may start with pre-conceived goals and embedded values in a given scenario, even if it was not the designer's intention (Poel 2020). Aggregation in this case is likely to require adjusting behaviour to accommodate another agent's values in a way that the original agent does not lose its purpose, rather than having it replicate a different set of values completely.

The risk of disempowering certain values mentioned in Section 4.1.1 is relevant here. Proper alignment should not just involve one agent's value system being subsumed by another's. It is not as simple as saying that autonomous agents should align with humans rather than the other way around, as autonomous agents are almost certainly acting as a proxy for another human or group of humans. A careful approach to understanding the context behind potential value disempowerment in aggregation, with appropriate compromises as a result, is necessary for ethical aggregation. Further examination of value compromise in other domains can lend insight here.

Disempowerment of values in value alignment may occur through combative AI whose goals differ from human stakeholders, potentially in the form of multiple competing autonomous agents acting as proxies for humans. This would lead to competition in attempts to achieve different world states that satisfy different values (Atkinson and Bench-Capon 2016; Nikolaidis et al. 2017). This is an example of what happens when alignment fails (Chaturvedi et al. 2023). As Hadfield-Menell and Hadfield (2019) described it, agents pursuing different rewards become engaged in a strategic game against each other to try and achieve their own goals, even if it reduces the gains for the other agent, as in the classic prisoner's dilemma (Rapoport and Chammah 1965).

The main form of combative AI we observed in our analysis, apart from discussion around manipulative superintelligences (Bostrom 2017; Li 2021; Murray 2017), was the case of autonomous agents overriding human decision-making. While the notion of surrendering our autonomy to autonomous agents may seem undesirable, research by Milli et al. (2017) suggested that there may be a level of nuance to this. In the case of humans

expressing orders imperfectly, the study by Milli et al. demonstrated superior performance of autonomous agents that attempted to support the human's underlying preferences. The research suggests that actions that may be seen as combative may have a time and place, but this still requires successful identification of a stakeholders' underlying values and careful ethical consideration of whether overriding is appropriate.

Transparency around the process of aggregation will also be essential. Given the need for value-based decisions about aggregation methods, different approaches may lead to different versions of alignment for the system, and different requirements to achieve alignment. Transparency will be necessary to mitigate the political risk this presents from empowered stakeholders like designers and owners having their values overrepresented compared to less powerful stakeholders like users. Given the need to adequately model stakeholder values in value-aligned system, this could present a novel opportunity to model what happens to these values during different aggregation mechanisms more explicitly. Both developing methods for aggregation, and understanding how they impact different stakeholders, would be an intriguing line of research.

Concluding this section on values, we have drawn attention to the key factors when trying to use values to promote aligned decision-making between humans and autonomous agents. It is important to understand that values are not just static objectives that can be identified once and forgotten about, but dynamic mechanisms for approval or disapproval of a situation, obscured by their latent nature and the complex interactions between values and context. Furthermore, the need to aggregate values for satisfying groups of stakeholders creates methodological and political opportunities and challenges for research. There is still a great deal of work to be done in modelling values properly in autonomous systems.

4.4 The Value Alignment Process

So far, we have discussed the themes that illustrate what is driving value alignment as a relevant field of research, the particular challenges it faces, and the mechanics of values, the field's core concept. We now focus on how the literature portrays the process of value alignment in systems. Our themes of *cognitive processes* and *human-agent teaming* together illustrate how value alignment occurs at the systematic level.

Analysis of the literature and our extracted codes led to the development of a concept for how the subprocesses of value alignment interact in the operation of a system. We split the subprocesses into two stages: *value identification and operationalisation* and *value calibration*.

Identification and operationalisation is the process of value communication and negotiation, then making those values functional in the relevant contexts for the sake of decision-making. We group these two verbs into one grouping of processes as the two are intertwined: operationalisation cannot occur without identification, but operationalisation can also reveal how values have not been sufficiently expressed or understood.

Value calibration accounts for the dynamic nature of values that we explored in Section 4.3.3. Its processes enable the communication of misalignment between agents and promote actions to correct it. Without calibration, value-aligned systems will inevitably become misaligned.

Although a system's life cycle will almost certainly start in identification and operationalisation, value alignment can move back and forth between the two groups of processes. Much like how operationalisation can state that values have not been understood sufficiently, calibration can reveal neglected values or stakeholders, or the need to redesign a system to attain changed values. Once the system has evolved to reflect this, calibration can continue. Hence, the process of value alignment between humans and AI agents is iterative.

4.4.1 Value Identification and Operationalisation. Value identification and operationalisation includes the processes of: value expression (including goal and preference expression); value aggregation; contextualisation; and value-based decision-making. We have already discussed expression in Section 4.2 and aggregation and contextualisation in Section 4.3, so here we focus on implementing values in relation to cognitive processes and human-agent systems. Viewing identification and operationalisation as part of a joint process combines the

technical and normative aspects of alignment mentioned in Section 4.1.2, by considering the means of selecting and implementing stakeholder values alongside the technical requirements in doing so. It also illustrates the close relationship between identifying the values of interest in a given context and then contextualising them to make them actionable by autonomous agents within the same context.

Value Learning, Teaching and Reasoning. Expressing our priorities, particularly in a dynamic system where the present agents may change over time, implies a need for other agents to be able to learn these priorities. Within our literature survey, the most common source of value-learning for autonomous agents was humans. This primarily seemed to occur due to the popularity of cooperative inverse reinforcement learning (Hadfield-Menell, Dragan, et al. 2016) as a value alignment mechanism, and indeed the method has evolved into an approach designated reinforcement learning from human feedback. A particularly modern example of this approach has been in the development of large language models, which are employed in a vast range of contexts with minimal oversight, in order to train these models to develop behaviour in line with our preferences (Bai, Jones, et al. 2022).

While humans make a natural candidate for teaching autonomous agents human values, challenges exist. Humans are cognitively limited in some ways compared to autonomous agents, given our need for rest and our restricted ability to interpret certain types of information compared to autonomous agents. As we mentioned in Section 4.2.1, expressing goals is difficult for humans, which complicates teaching them to autonomous agents. This is accounted for by some modern approaches, however, which aim to learn human preferences as these are easier for humans to express (Christiano et al. 2017; Zintgraf et al. 2018). Regardless, our cognitive limitations remain an important consideration in any human-driven teaching process.

Discussion drew a distinction between two types of reasoning: theoretical reasoning, which concerns what is the case; and practical reasoning, which concerns what to do (Zurek, Araszkiewicz, et al. 2022). When it came to interpreting value-based knowledge learned from humans, practical reasoning was considered by the papers in our survey more than theoretical reasoning.

While discussion focused more on practical reasoning, theoretical reasoning becomes very important if we consider the suggestion in Gabriel (2020) that "AI would have to be aligned with some set of beliefs about value, not with value itself.". This would suggest that, for successful alignment to occur, autonomous agents are able to evaluate their beliefs about how values are realised, incorporate uncertainty into the reasoning process, and determine if these beliefs are consistent with humans in the system in which they exist. We can link this to human decision-making, as this was another process used to consider autonomous agent decision-making. A prominent observation was that humans make ethical decisions with incomplete and imperfect information (Cervantes, Rodríguez, et al. 2016; Robinson 2023), and it is natural to assume that autonomous agents will have to make ethical decisions under the same circumstances. Bogosian (2017) and Russell (2019) and Eckersley (2019) all supported the necessity of implementing uncertainty in autonomous agents' moral decision-making to achieve alignment.

An interesting aspect of the reasoning process about values is the idea of *emotional intelligence* in AI. This includes both recognising emotions in humans as a form of feedback, and autonomous agents using their own form of emotions in their reasoning process. While Constantinescu and Crisp (2022) observes that emotionless AI would be advantaged in making decisions free from undesirable emotions, it would also limit their ability to act in a virtuous capacity. Cervantes, Rodríguez, et al. (2016) also raises the need for emotions in effective decision-making, citing neuroscientific evidence of the role of emotions in social and ethical contexts, both highly relevant in value alignment.

Decision-Making with Values. Given our tendency to anthropomorphise autonomous agents (Duffy 2003; Salles et al. 2020), it would seem natural that discussion of human decision-making would appear in works examining decision-making frameworks for autonomous agents. Butlin (2021) suggests that there is an acceptance that humans and animals use multiple decision-making systems, and this includes both a model-free and model-based

reinforcement learning style approach. As Cervantes, Rodríguez, et al. (2016) observes, human decision-making happens in a continuous fashion, with the majority of it being unconscious. In the context of values, Zurek and Mokkas (2021) mentions that humans do not compare individual values and how they will be promoted but evaluate various values simultaneously and use this to evaluate options. While emphasis in value alignment tends to be on constructing effective decision-making for autonomous agents, understanding how humans make decisions is essential in considering how both types of agents will interact with each other in pursuing alignment.

While no perfect formalism for how humans reason with values exists, it can generally be understood that values impact decision-making through contextualisation of values and the internal ordering over the values (Liscio, Lera-Leri, et al. 2023; Serramia, Lopez-Sanchez, et al. 2020), in order to evaluate and choose from the available options (Szabo, Such, and Criado 2020). As Han et al. (2022) states, humans feel a "calling" to realise positive values, an "ought to do". But as Waser (2015) points out, while these values have enabled our survival to modern times, a lack of explanatory power regarding how values function makes their use difficult in ethical dilemmas without clear value resolutions. Simply having agents learn latent human values and mimic them in decision-making may not be ideal for complex ethical dilemmas, if such dilemmas lack consensus on a reasonable outcome. Indeed, this lack of explanatory power connecting values to decision-making increases the risk associated with relying on autonomous agents learning our values and trying to emulate them, as their behaviour becomes less predictable as a result, as discussed in Section 4.1.1. However, given that by definition we want autonomous agents to do the things we value, trying to remove values from the process of integrating AI agents into systems would seem self-defeating.

Naturally, as autonomous agents make more decisions, we can expect them to face more ethical dilemmas as well. As Gabriel (2020) discusses, there is an interpersonal dimension to these conflicts: what is value-aligned for one person won't necessarily be for others. This brings us back to the challenges of value aggregation raised in Section 4.3.4, and the associated challenges with diverse perspectives. Given the perpetual conflict between humans' values and our construction of ethical frameworks that attempt to resolve these dilemmas, the choice of conflict resolution approaches will inevitably embody at least one ethical framework. Cervantes, López, et al. (2020) cites Broeders et al. (2011) in how individuals use the moral rule that has given them the most success in the past when making judgments in similar situations, which Cervantes, López, et al. (2020) incorporates into their own framework. From a design decision, it may be practical to focus on developing autonomous agents to recognise these ethical dilemmas and defer to humans, rather than rely on them to resolve them for us, if we wish to maintain agency over our morality. Empirically, we may need to experiment with autonomous agents using different dilemma resolution methods in different contexts to determine which will be palatable in practice, as theorising alone seems unlikely to lead to a resolution.

For autonomous agent decision-making, utility-based approaches to learning values and ethics were abundant in our survey. However, little reasoning was done over the learned utility functions beyond taking the action that maximises the utility score. While this is the norm in utilitarian decision-making, critics point out that such an approach can be disastrous when the utility score fails to accurately capture the underlying objective (Gabriel 2020; Zhuang and Hadfield-Menell 2020). Murray (2017) proposes that such systems should integrate a form of temperance into their reasoning, subjecting their maximisation goals to constraints on their impact. This may also have useful implications for aggregating values between stakeholders, by incorporating agreed trade-offs between groups while still achieving satisfying outcomes.

Other authors investigated utility with multiple dimensions. This emulates value pluralism (Mason 2006), the assumption that multiple, incommensurable values exist. As stated in Szabo, Such, Criado, and Modgil (2022), this assumption of multiple values existing is standard in value-based reasoning, either due to the distinctness of values, or the impracticalities of attempting to collapse them into a single value. Some works that included this multi-objective approach to utility functions include Haas (2020), Peschl et al. (2022), Rodriguez-Soto, Lopez-Sanchez, et al. (2021), and Vamplew et al. (2018).

Some other examples of value-based reasoning for decision-making in the survey were Atkinson and Bench-Capon (2016), who demonstrated how value based argumentation could be used to select actions without assuming other agents' preferences; Holgado-Sánchez et al. (2023), who examined how different contextualisations of values would impact the range of admissible actions; and Zurek, Araszkiewicz, et al. (2022), who used principle-based reasoning to select goals based on the values to be promoted.

4.4.2 Value Calibration. Value calibration is centred around evaluation, feedback, and adjustment. Proposed previously by Firt (2023) in the dimension of trust, we extend this stage to be necessary to accommodate the dynamic nature of values and context, as well as to allow corrections for the inevitable noise generated by the difficulties in value expression, learning and aggregation.

Feedback. Throughout the learning processes observed in the survey, the concept of feedback was pervasive. Given the difficulty in expressing values we discussed in Section 4.2.1, and the dynamism of values discussed in Section 4.3.3, it seems sensible to believe that by attempting to align to a set of values at one point in time, we may fail to confirm that alignment is still maintained when we assess again at a later point. This could be due to misrepresentation of the values in the first place, or changes in the target values over time. To accommodate this, a capability for an autonomous agent to adjust its values and priorities at a given moment in response to feedback, potentially needing to learn new value representations in the process, will be required for robust value-aligned systems. This could be considered one of the key problems facing value alignment.

In particular, when human teachers are engaged with the agent in providing this feedback, the teaching process also needs to support the human teacher's understanding of how their feedback has been interpreted by the autonomous agent. Sanneman and Shah (2023a) explores this area through the use of explainable AI. Humans are also fallible and may make teaching errors (Milli et al. 2017) or teach undesirable values (relative to other stakeholders' values) to an autonomous agent (Gabriel 2020). These issues reflect the need for transparency and explainability in the human-teaching process, both the values that have been learnt by the autonomous agent, for auditing and to support the human teacher, and the values held by human teacher, for the sake of auditing what values a system has been trained on.

Also relevant to the topic of feedback is the role of self-reflection in autonomous agents, as a form of internally-generated feedback. Murray (2017) discusses this in the context of stoic ethics, where self-reflection would be used by autonomous agents to evaluate whether its past behaviour was aligned with its desired behaviour, while regret could be used in cases with no good options, to drive the agent to seek alternatives in similar future scenarios. Shaw et al. (2018) considers self-reflection in a more normative sense, where it could be used to flag inconsistencies in an autonomous agent's learned principles. Stenseke (2023) also proposes a self-reflection framework using a reflective and proactive method. The reflective method evaluates past behaviour similar to the work in Murray (2017), while the proactive approach allows the agent to simulate future scenarios to evaluate its potential behaviour. While self-reflection as a mechanic is still largely conceptual, it offers exciting potential for the value alignment process in the face of shifting and obscured alignment needs.

The idea of emotional intelligence in autonomous agents is again relevant here, as emotions can serve as a form of feedback that can signal a need for an agent to adjust its behaviour (Martınez-Miranda and Aldea 2005; Salloum et al. 2025). Emotions have been suggested as useful for inferring non-observable mental states (Tzeng 2022), which can be used for triggering responses by the agent (Harland et al. 2023), as well as for acting believably in human-computer interactions (De Carolis et al. 2010) while avoiding acting in a manipulative way (Murray 2017; Scheutz 2011). Han et al. (2022) also suggests that a need to feel emotions would be required for an autonomous agent to become independently value-aligned in a material value-ethics paradigm, where knowledge of values is only conveyed through emotions. Given their ability to contribute to reasoning about and learning values, the study of emotions in autonomous agents is an intriguing avenue for value alignment work.

Evaluation. We round out our analysis with methods of evaluating whether systems have actually achieved value alignment or not. Given the challenges in modelling values that we have outlined up to this point, this is naturally a difficult assessment to make.

Our analysis agreed with the observation in Feffer et al. (2023) that most evaluations in the research space rely on simulations or theoretical proofs. We were concerned by the lack of uniformity in the scenarios used for testing: authors would usually create novel simulations or repurpose their earlier work when testing their methodologies. Absent from the literature were reliable baseline scenarios which would support the comparison of different approaches to value alignment.

While the trolley problem, a well-explored ethical dilemma example in moral philosophy (Thomson 1984), was discussed in some studies (Faulhaber et al. 2019; Loreggia et al. 2019; Peterson 2019), Eckersley (2019) is quick to point out the negligible relevance of the trolley problem to the space of possible scenarios an agent might have to consider in value alignment. We have illustrated other significant challenges in value alignment, such as goal expression and context modelling, to which the trolley problem adds little insight. Our conclusion is that the trolley problem could prove a useful baseline for developing theoretical reasoning about values in autonomous agents, given that its well-studied nature provides complimentary thought on any produced reasoning. However, authors should be wary of confirmation bias when using the trolley problem and ideally test their model on alternative value-based scenarios.

We theorise that there are two causes for the lack of baseline scenarios in the surveyed literature. First, ambiguity on what value alignment actually is makes it difficult for authors to assess whether scenarios from other research would be suited for their own interpretation of value alignment. The goal of this paper is to help alleviate this by working towards unifying the nature of the value alignment topic, while also identifying sub-problems for which tractable baselines could be generated. The ability to reproduce a human's identified values through agent interaction, would be one example baseline. Successful context modelling, and hence policy switching, could be another - one we already see studied through explainable AI approaches to domains like self-driving cars (Atakishiyev et al. 2024).

Second, it is difficult to envision scenarios where different values and goals can be clearly articulated, simple enough to implement and test repeatedly, while also avoiding overly simple behaviour that has an "obviously correct" answer. For example, Rodriguez-Soto, Serramia, et al. (2022) assesses their value alignment methodology through their public civility game, where an agent must learn to throw a rubbish bag towards the rubbish bin, rather than into its fellow agent's path. While this is fine for confirming that their agent has learned desirable behaviour, there is no real alternative behaviour here that could also be considered aligned from a different agent's perspective. Someone may be in favour of throwing trash in their colleague's path but attempts to justify this would likely introduce new values such as this being the more efficient option or simply preferring to be mean-spirited. There is a difficult trade-off in designing scenarios between the necessary complexity of behaviour and ease of implementation of test scenarios: additional complexity increases the difficulty in constructing such a scenario, as well as measuring the effect of different factors on outcomes, but also can lead to more valid evaluation of value-based behaviour. The development of these baseline test scenarios, and the demonstration of their fitness for purpose, would be a valuable contribution to value alignment research.

Assessing value alignment is not a one-off exercise but rather an ongoing process. As we have established in Section 4.3.3, values change over time through contexts and stakeholders changing. And as Poel (2020) points out, the evolution of systems of autonomous agents can lead to the embedding or dis-embedding of different values over time as the agents adapt within the system. Questions regarding the frequency of this assessment in the face of dynamic values, and how to respond when misalignment is detected, remain open.

Finally, we emphasise the point made previously by Sanneman and Shah (2023b) that there is a significant shortage of empirical testing of value alignment systems, including both humans and autonomous agents. Simulations and theoretical approaches are useful for testing methodologies or ideas initially. However, our

interest in value alignment comes from the need to have humans and autonomous agents interact, and we have established that values in practice are highly dynamic and interactions with them are difficult to predict. This is made more difficult by the cognitive hurdles in interpreting values and validating that alignment has happened, particularly in less controlled environments than what experiments usually provide. Research requires contextual interaction with human stakeholders to generate data that can be placed in the context of the system's intended operating environment, and the values and contextualisations these imply. Unfortunately, this is difficult to implement for many systems, given that the risk of being misaligned with stakeholder values in many scenarios can cause harm, be it physical, mental, or otherwise.

While studies making use of humans were encountered in our survey (Ficici et al. 2008; Liscio, Meer, et al. 2022; Siebert et al. 2022; Svegliato et al. 2020; Zintgraf et al. 2018), the focus of these studies was on modelling user values and preferences. This is an essential step to value alignment, but it only approaches the value identification stage of the process. Assessment of a system's alignment with a stakeholder's values will require some ability to assess system alignment in its operating contexts, and a means to validate this alignment for auditors. This brings us back to the need for formalised mechanisms for evaluating alignment.

Two studies did investigate humans exploring components of the system or interacting with the system to assess alignment. Sanneman and Shah (2023b) used latent factor analysis to investigate alignment between stakeholder goals and agent reward function, identifying two latent measures they labelled 'feature alignment' and 'policy alignment'. Relating these measures to our analysis, we would describe them respectively as the need to represent values appropriately for stakeholders and prioritise different values via proxies in an agreeable manner. Nikolaidis et al. (2017) investigated the performance of human-robot teams cooperating in situations where the robot could override human preferences. It was observed that an adaptable approach to the human's degree of cooperation improved performance relative to pure obedience. While the lens of this study was on user trust rather than value alignment, it still provides insight into the nature of humans interacting with autonomous agents and how their goals, which are derived from values, can be supported through the nature of said interaction.

While theory development and simulations are essential tools in the development of value-aligned systems, studies like the previous two demonstrate the benefits and needs for more empirical research. Given the subjective nature of values and the relatively limited experience of humans interacting with autonomous agents, involving humans in more stages of both researching and developing value-aligned systems is vital to advance our understanding about the possibilities and limitations of value alignment and how these can be achieved.

In conclusion, we see that the value alignment process is a dynamic human-AI process, and a hugely complex one at that. Achieving alignment between humans and autonomous agents in systems will require more than just teaching autonomous agents to learn a set of values: it will require understanding how these values are communicated and modelled by both agent types and how feedback can be used to adjust them as necessary. In addition to the essential need to model context appropriately, other forms of information like emotion can provide insight in the alignment process, Finally, understanding how the interactions between both humans and autonomous agents impacts alignment and assessing the state of alignment in the face of these ongoing interactions is not currently well understood, but this needs to change for robust alignment to be achieved.

5 Discussion

Our objective for this paper was to gain a deeper understanding of the value alignment challenge for humans and autonomous agents. From our analysis we can define value alignment as an ongoing dynamic process of identifying, operationalising and calibrating values, that is complicated by the abstract nature of values and contextualization, the difficulties in identifying and communicating values between humans and autonomous agents, evaluating the state of values, accommodating the dynamic nature of values, and the ethical and political risks that design decisions around values and their aggregation entails.

From our analysis, it is clear that value alignment is a complex topic. It cannot be described in terms of simple objectives, as the concepts of 'value' and 'alignment' are themselves inherently complex and subject to interpretation. The scope of processes involved in value alignment and the range of subjects that can contribute towards understanding it is broad. Effectively formalising value alignment as a process that can be implemented reliably will require a concerted interdisciplinary effort to reach viable solutions and should not be viewed only as a computer science concern.

Furthermore, researchers will benefit from narrowing their focus to particular aspects of the value alignment process, such as value representation or managing uncertainty in environmental perception. In doing so, they should not lose sight of the wider elements of value alignment and how their efforts may impact upstream or downstream processes, given the ongoing nature of value alignment as a process.

Value alignment should also be understood as a process of human-agent interaction, as we see through the need to calibrate alignment through teaching values and feedback between humans and autonomous agents. Attempting to approach value alignment while focusing only on AI mechanisms, under the assumption that humans can be integrated later, risks starting with a false premise. Ignoring the humans, the sources of values and the drivers of their dynamic and complex nature, in the process of value-alignment is difficult if not outright naive.

We emphasise that value alignment is an iterative process. We have made clear the dynamic nature of values, being highly sensitive to context and stakeholders, and these will change throughout the system's operation. This makes alignment inherently unstable, and even if the first version of an agent deployed and evaluated appears appropriately aligned, this is no guarantee of future alignment. This iterative nature emphasises the role of online learning and self-maintenance of autonomous agents in value alignment contexts.

Regular evaluation and adaptability are essential to maintain robust value-aligned systems. Interaction between humans and autonomous agents to produce feedback is also key to reaching an aligned state, through communicating states of misalignment and providing guidance to correct it. Not only will this enable alignment, but interaction also has scope for improving user trust towards such systems and supporting the appropriate adoption of autonomous agents in society.

It is clear from our analysis that value-aligned systems need more empirical research if the field wants to continue advancing. There are numerous challenges in the successful expression of human values across diverse contexts, as well as too many uncertainties in how humans will respond to attempts to embed these values in autonomous agents or the impact these agents will have on human values. These cannot be anticipated purely from theoretical or simulation approaches alone.

This should not be taken as an indication that theory and simulation have no use, as designing autonomous systems is a lengthy and costly process, and attempting to start with humans involved will inevitably add hurdles in undertaking the research. We instead believe that we are in a position now where empirical data on the value alignment process can add value, and researchers should embrace it. Collecting this empirical data can naturally invite risks when exposing humans to autonomous agents that may be misaligned with them, so care must be a priority. Researchers with experience in human trials can easily lend expertise here.

The fact that value alignment is difficult should not be understated. It is very doubtful that a single mechanism can overcome the numerous challenges presented by value alignment. Instead, a combination of components external and internal to the autonomous agent undergoing alignment will likely be needed. Overall alignment will also require regular monitoring during the system's lifetime to promote calibration, and alignment will also require mechanisms to enable coordination between both human and autonomous agents. With this in mind, it is important to build resilience into a system for when misalignment inevitably does occur among these moving parts.

Finally, it is important to note that alignment should not be expected to be achieved with all potential stakeholders. However, in some circumstances misalignment can be acceptable, or at least tolerable. A generative

art model that produces inoffensive art that is not to some users' tastes would not be aligned with them, but this is unlikely to be controversial, and alternatives may exist that can replace the given model. In comparison, a system that prejudices a certain group towards more severe jail sentences would be misaligned with that group, and this would not be acceptable as the harm to their values of justice and freedom would be severe.. Deciding when misalignment is acceptable is itself a value-laden decision, and another part of the complex process of designing value-aligned systems.

5.1 Opportunities and Future Directions

Given its nature as a complex, interdisciplinary topic, there is much value to be gained from other disciplines' insights into the core challenges we have identified in this paper. Table 2 illustrates the subjects encountered in this analysis, but it is not exhaustive with regard to what disciplines can contribute. For example, human-computer interaction as a research topic could be critical in the calibration group of processes, as the field could offer insight into how to engage stakeholders interacting with autonomous agents to assess whether they feel the system is aligned or not, or develop better methods for them to use in reporting their values to autonomous agents. Reducing barriers for interdisciplinary knowledge sharing and collaboration in value alignment research will be very useful.

Tackling one of the main challenges in value alignment, a better understanding of how values and goals are expressed, and where mistakes in this process occur, would support alignment. Given the cognitive difficulties involved in expressing these concepts, attempts to formalise some processes in the context of aligning autonomous agents could potentially control for these challenges and reduce errors in identification.

Alternatives to utility function, which as discussed in Section 4.2.1 are prone to under-specifying objectives or integrating more refined forms of utility functions into value alignment, are also due for investigation. Other approaches to modelling values entirely, such as attempts to encode virtues or deontic systems, may be necessary for achieving alignment, but these are still works in progress.

Value aggregation is also worthy of research. The current standard in study involves aligning two agents, usually at least one of which has a static policy. Pushing for research that attempts to align more than two agents using utility functions would be a relatively low-effort but simple starting challenge, especially given the existing work on using norms for multi-agent systems, and some of this research no doubt already exists in the multi-agent or swarm AI space. Potentially of greater interest would be examining different methods of alignment and their impacts on the form alignment takes, particularly in what stakeholders are disadvantaged or empowered. Such results would not only be useful in supporting design decisions but also in understanding the political implications of autonomous agents as a technology and how these can be accounted for.

The contextualisation of values as a process is currently vague, but it is critical to value alignment. Understanding and formalising this process, even if it is only in the context of aligning autonomous agents, could add immense value to the topic and improve transparency in design. Case studies of how values have already been embedded in autonomous agents through processes like value-sensitive design (Friedman 1996) and how these relate to value alignment would be particularly useful in generating empirical results.

Core to modelling contextualisation is the need for a means to model context and how it is recognised by autonomous agents. This could support the explainability of agent decision-making by indicating the state features that describe a given context, possibly giving a compressed representation of all state features. If this can be further enhanced by understanding how contexts affect values specifically, then this would further improve the transparency of decision-making in autonomous agents. This is no straightforward task, but it would add necessary value in supporting value alignment, so worth our continued attention.

Value calibration was neglected in the research we analysed, but given the dynamic nature of values, this needs to change. The need for a means to effectively track stakeholders' values in dynamic situations is a component

of this. If it quickly becomes apparent on deployment that misalignment is an issue, this needs to be identified before negative outcomes occur. This links into the need for understanding how often evaluation and calibration need to occur, which itself points to a need for a better understanding of value dynamism. In addition, further development of the methodologies for engaging stakeholders and assessing their alignment with active systems will be critical to effective calibration.

Testing approaches to value alignment is one last area that would strongly benefit from attention and unification. The current environment of individual experimental designs, often lacking in complexity or validation, limits the ability to say whether an approach is fit for purpose as it prevents comparison with other methods and obscures the complexity of values and context. Resolving this will require focused attention on what makes a good value alignment test scenario, and to what extent these can be generalised between environments to produce benchmarks. This is also another area which invites empirical research, and expertise from those with experience in measuring human-computer interaction.

A challenge with any thorough systematic review is the time required to execute it properly. Our survey covers publications up to the end of 2023, based on the material we retrieved at our search date. However, value alignment is a fast-moving field, as demonstrated in Fig. 2. Since our search date, there has been an explosion of interest in alignment applied to large language models (Bao et al. 2025a; Huang, Liu, Guo, T. Sun, J. Sun, Wang, Zhou, Wang, Teng, Qiu, et al. 2023; Kwon et al. 2024; Padhi et al. 2024), as well as novel alignment approaches such as alignment faking (Chaudhury and Shiromani 2025), agentic misalignment (Kierans et al. 2025) and constitutional AI (Bai, Kadavath, et al. 2022).

These developments do not appear to undermine our analysis but instead are consistent with our identified problems. In the large language model value alignment literature, we see many of the topics raised in this survey appearing. Dognin et al. (2025) discusses value aggregation, contextual values and value modelling by developing a system to aggregate language models trained on particular values, and weigh their responses based on the user's given value profile, and Padhi et al. (2024) also touches on contextual values by identifying the need for language models to be able to adapt readily to different value systems. The idea of value calibration was also referenced, both through calibrating the users' values through interaction with an LLM (Gomez Tobon and Law 2025) and how prompt attacks could be used to de-calibrate an LLM, and how to prevent this (Bao et al. 2025b). The topic of value identification, particularly understanding the values learned by the LLM, was also discussed (Huang, Liu, Guo, T. Sun, J. Sun, Wang, Zhou, Wang, Teng, Qiu, et al. 2024; Xu et al. 2024; Ye et al. 2025). We also see work in value awareness engineering (Holgado-Sánchez 2023; Osman and d'Inverno 2024) emerging, a modern paradigm for value alignment that also recognises the contextual, dynamic nature of values and the need for an interdisciplinary approach to the topic.

Our review frames value alignment independently of these recent developments. That said, a deeper analysis of value alignment papers published after 2023, replicating the methodology in this paper, would be a valuable follow-up work. This new survey, in tandem with our current analysis, would be an excellent way to compare value alignment research before and after the widespread adoption of LLMs.

5.2 Limitations

As stated in Section. 3, we limited the content of our survey to the English-language papers due to a lack of translation capability. We also observed that non-Western ethical systems and values were neglected in the value alignment research that we analysed. Given that this paper was authored by a Western team, this would have further compounded the effect of Western values on the interpretation of the data.

As a result, this paper presents a dominantly Western perspective on the process of value alignment. This is not ideal, as value alignment will be required in technologies around the globe, and autonomous agents will need to process values in numerous regions, not just the West. Future studies should keep this in mind and try to

bridge this gap. Complementing our understanding of value alignment with further research on non-Western approaches could offer insights into geographic differences in the value alignment process and challenges that may arise in the face of globalised AI technology.

Another limitation of our analysis is our use of peer-reviewed academic literature from Scopus to source our data. Scopus does not include content from most workshops and a significant number of conferences in computer science. While our bibliography search extended our paper list, it still relied on relevant papers being cited in Scopus-sourced papers to begin with. Extending our study with additional literature databases would be a valuable addition.

While we analysed academic literature, value alignment is also a popular topic of discussion on non-academic platforms frequented by industry practitioners and non-academic thinkers, and there is a significant body of writing on these platforms. This content is potentially not subject to the same scrutiny as academic literature, which is why we did not include it in our review, but it may still offer insight in future analysis. Even if it does not provide any new insights into the value alignment process, it would give an insight into the thought processes of the people doing value alignment in many modern AI systems.

6 Conclusion

The goal of this paper was to review different perspectives in the literature in order to understand value alignment through thematic analysis. Our research has led us to define value alignment as a complex, interdisciplinary topic with an interconnected set of themes. These themes lead us to the idea that value alignment is about enabling humans and autonomous agents to interact in ways that support multiple competing dynamic human values, which are highly sensitive to the operating context and constrained by expression challenges. The numerous challenges emerge from how these values are identified and operationalised in ways compatible with both humans and autonomous agents across multiple contexts, and then how alignment in terms of these values is calibrated throughout the system's lifetime in the face of political risks and ethical disagreement.

In summary, these are our key observations from this survey:

- Value alignment is *complex*. Value alignment as a process cannot be described in terms of simple objectives. The scope of processes involved in value alignment, and the range of subjects that can contribute towards understanding it, is broad. The process will require a concerted interdisciplinary effort to reach viable solutions and should not be viewed only as a computer science concern. Researchers will also benefit on recognising the sub-problems in value alignment requiring attention, such as value modelling or alignment evaluation, rather than viewing value alignment as a single monolithic problem.
- Value alignment is a process of *human-machine interaction*. Attempting to approach it while focusing only on the technical or normative dimensions risks starting with a false premise, due to the interconnected nature of the social and technical dimensions in developing value-aligned systems. Successful research and development of value-aligned systems needs to move beyond theory and simulations to include more empirical research involving humans, or at least an understanding of human-agent interaction, if the field wants to continue advancing.
- Value alignment is *iterative*. Values are highly sensitive to context, and operating contexts will change repeatedly throughout the system's lifetime. Even if the first version of an agent deployed is appropriately aligned, this will not last forever. Adaptability is essential to maintain robust value-aligned systems. Interaction between humans and autonomous agents is also key to reaching an aligned state, through communicating states of misalignment and providing guidance to correct the agents' behaviour. Not only will this enhance alignment, but this interaction will also improve user trust towards the system and support the adoption of autonomous agents by society.

- Value alignment is *two-way*. Humans and autonomous agents will adapt to each other. While autonomous agents will ideally learn to embody human values, there are risks around humans deviating from their own ideal values through interaction with autonomous agents. To avoid this leading to undesirable value changes, value checks should be available for both humans and autonomous agents engaging with a system.
- Value alignment is *difficult*: the abstract and dynamic nature of values combined with the sensitivity of the contextualisation process makes value alignment a very unstable process. It is very doubtful that a single mechanism can resolve the diverse issues in value alignment; instead, a combination of components external and internal to the autonomous agent undergoing alignment will be needed, and overall alignment will require regular monitoring during the system's lifetime. With this in mind, it is important to build resilience into a system for when misalignment inevitably does occur.

As autonomous agents become more embedded in our society, effective value alignment will be crucial to ensure these agents behave in ways that support our goals rather than oppose them, and this includes robust methods for assessing when misalignment is occurring for each of the system's stakeholders. By drawing attention to the many aspects of this process, we hope that this paper will support future researchers wanting to engage in the field by clarifying the research opportunities available to them, as well as flagging potential risks that should be considered in developing value-aligned systems.

Acknowledgments

This work is supported by the UK Research and Innovation under Grant No.: EP/S023437/1.

Open Source Statement

The data used in this paper is publicly available at https://github.com/JamMack/Understanding-Value-Alignment-as-a-Process-a-Survey?tab=readme-ov-file.

References

P. Alamdari, T. Klassen, R. Icarte, and S. McIlraith. 2022. "Be Considerate: Avoiding Negative Side Effects in Reinforcement Learning". In: *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS*. Vol. 1, 18–26. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85134303258&partnerID=40&md5=033693d752e91cfbf3db87d7aec0399b.

A. Aler Tubella, A. Theodorou, F. Dignum, and V. Dignum. 2019. "Governance by glass-box: Implementing transparent moral bounds for AI behaviour". In: *IJCAI International Joint Conference on Artificial Intelligence*. Vol. 2019-August, 5787–5793. doi:10.24963/ijcai.2019/802.

A. Aler Tubella, A. Theodorou, V. Dignum, and F. Dignum. 2019. "Governance by glass-box: Implementing transparent moral bounds for AI behaviour". *arXiv preprint arXiv:1905.04994*.

N.-M. Aliman and L. Kester. 2019. *Augmented utilitarianism for AGI safety*. Vol. 11654 LNAI. Pages: 21. doi:10.1007/978-3-030-27005-6_2.

N.-M. Aliman, L. Kester, P. Werkhoven, and R. Yampolskiy. 2019. *Orthogonality-based disentanglement of responsibilities for ethical intelligent systems*. Vol. 11654 LNAI. Pages: 31. doi:10.1007/978-3-030-27005-6_3.

C. Allen, I. Smit, and W. Wallach. Sept. 2005. "Artificial Morality: Top-down, Bottom-up, and Hybrid Approaches". en. *Ethics and Information Technology*, 7, 3, (Sept. 2005), 149–155. doi:10.1007/s10676-006-0004-4.

T. Arnold, D. Kasenberg, and M. Scheutz. 2017. "Value Alignment or Misalignment – What Will Keep Systems Accountable?" en.

K. J. Arrow. 2012. *Social choice and individual values*. Vol. 12. Yale university press.

S. Atakishiyev, M. Salameh, H. Yao, and R. Goebel. 2024. "Explainable artificial intelligence for autonomous driving: A comprehensive overview and field guide for future research directions". *IEEE Access*.

K. Atkinson and T. Bench-Capon. 2016. *Value based reasoning and the actions of others*. Vol. 285. Pages: 688. doi:10.3233/978-1-61499-672-9-680.

C. Badea and G. Artus. 2022. *Morality, Machines, and the Interpretation Problem: A Value-based, Wittgensteinian Approach to Building Moral Agents*. Vol. 13652 LNAI. Pages: 137. doi:10.1007/978-3-031-21441-7_9.

Y. Bai, A. Jones, et al.. 2022. "Training a helpful and harmless assistant with reinforcement learning from human feedback". *arXiv preprint arXiv:2204.05862*.

Y. Bai, S. Kadavath, et al.. 2022. "Constitutional ai: Harmlessness from ai feedback". *arXiv preprint arXiv:2212.08073*.

Z. Bao, Y. Ji, W. Wu, X. Chen, and L. He. 2025a. "Supervisor Alignment Framework: Enhancing LLM Alignment with Query-Ignoring Strategy and Multi-Agent Interaction". In: *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

Z. Bao, Y. Ji, W. Wu, X. Chen, and L. He. 2025b. "Supervisor Alignment Framework: Enhancing LLM Alignment with Query-Ignoring Strategy and Multi-Agent Interaction". In: *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. doi:10.1109/ICASSP49660.2025.10890479.

W. A. Bauer. Mar. 2020. "Virtuous vs. utilitarian artificial moral agents". en. *AI & SOCIETY*, 35, 1, (Mar. 2020), 263–271. doi:10.1007/s00146-018-0871-3.

D. Behdadi and C. Munthe. June 2020. "A Normative Approach to Artificial Moral Agency". en. *Minds and Machines*, 30, 2, (June 2020), 195–218. doi:10.1007/s11023-020-09525-8.

T. Bench-Capon and S. Modgil. 2017. "Norms and value based reasoning: justifying compliance and violation". *Artificial Intelligence and Law*, 25, 1, 29–64. doi:10.1007/s10506-017-9194-9.

T. Bench-Capon. 2020. "Ethical approaches and autonomous systems". *Artificial Intelligence*, 281. doi:10.1016/j.artint.2020.103239.

H. Bharadhwaj. 2021. "Auditing Robot Learning for Safety and Compliance during Deployment". In: *Proceedings of Machine Learning Research*. Vol. 164, 1801–1806. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85159450387&partnerID=40&md5=410df4b991b07bbff57d096928a03f3c.

K. Bogosian. 2017. "Implementation of Moral Uncertainty in Intelligent Machines". *Minds and Machines*, 27, 4, 591–608. doi:10.1007/s11023-017-9448-z.

N. Bostrom. 2017. *Superintelligence: paths, dangers, strategies*. Dunod.

V. Braun and V. Clarke. 2006. "Using thematic analysis in psychology". *Qualitative research in psychology*, 3, 2, 77–101.

R. Broeders, K. Van Den Bos, P. A. Müller, and J. Ham. 2011. "Should I save or should I not kill? How people solve moral dilemmas depends on which rule is most accessible". *Journal of Experimental Social Psychology*, 47, 5, 923–934.

D. Brown, J. Schneider, A. Dragan, and S. Niekum. 2021. "Value Alignment Verification". In: *Proceedings of Machine Learning Research*. Vol. 139, 1105–1115. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85158007676&partnerID=40&md5=ad685f88b92d8aea1c2525d3aa3b1231.

P. Butlin. 2021. "AI Alignment and Human Reward". In: *AIES 2021 - Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 437–445. doi:10.1145/3461702.3462570.

M. Cappuccio, E. Sandoval, O. Mubin, M. Obaid, and M. Velonaki. 2021. "Can Robots Make us Better Humans?: Virtuous Robotics and the Good Life with Artificial Agents". *International Journal of Social Robotics*, 13, 1, 7–22. doi:10.1007/s12369-020-00700-6.

M. L. Cappuccio, A. Peeters, and W. McDonald. Mar. 2020. "Sympathy for Dolores: Moral Consideration for Robots Based on Virtue and Recognition". en. *Philosophy & Technology*, 33, 1, (Mar. 2020), 9–31. doi:10.1007/s13347-019-0341-y.

J.-A. Cervantes, S. López, L.-F. Rodríguez, S. Cervantes, F. Cervantes, and F. Ramos. Apr. 2020. "Artificial Moral Agents: A Survey of the Current Status". en. *Science and Engineering Ethics*, 26, 2, (Apr. 2020), 501–532. doi:10.1007/s11948-019-00151-x.

J.-A. Cervantes, L.-F. Rodríguez, S. López, F. Ramos, and F. Robles. Apr. 2016. "Autonomous Agents and Ethical Decision-Making". en. *Cognitive Computation*, 8, 2, (Apr. 2016), 278–296. doi:10.1007/s12559-015-9362-8.

S. Chaturvedi, C. Patvardhan, and C. Lakshmi. 2023. "AI Value Alignment Problem: The Clear and Present Danger". In: *2023 6th International Conference on Information Systems and Computer Networks, ISCON 2023*. doi:10.1109/ISCON57294.2023.10112100.

A. Chaudhury and S. Shiromani. 2025. "ChameleonBench: Quantifying Alignment Faking in Large Language Models". In: *The 17th Asian Conference on Machine Learning (Conference Track)*.

P. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei. 2017. "Deep reinforcement learning from human preferences". In: *Advances in Neural Information Processing Systems*. Vol. 2017-December, 4300–4308. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85046999643&partnerID=40&md5=5ffef52e69ce75915865634de63042ba.

K. Coleman. 2001. "Android arete: Toward a virtue ethic for computational agents". *Ethics and Information Technology*, 3, 4, 247–265. doi:10.1023/A:1013805017161.

M. Constantinescu and R. Crisp. 2022. "Can Robotic AI Systems Be Virtuous and Why Does This Matter?" *International Journal of Social Robotics*, 14, 6, 1547–1557. doi:10.1007/s12369-022-00887-w.

M. Constantinescu, C. Voinea, R. Uszkai, and C. Vică. 2021. "Understanding responsibility in Responsible AI. Dianoetic virtues and the hard problem of context". *Ethics and Information Technology*, 23, 4, 803–814. doi:10.1007/s10676-021-09616-9.

P. Córdova and R. Vicari. 2022. *Practical Ethical Issues for Artificial Intelligence in Education*. Vol. 1720 CCIS. Pages: 445. doi:10.1007/978-3-031-22918-3_34.

S. Cranefield, M. Winikoff, V. Dignum, and F. Dignum. 2017. "No pizza for you: Value-based plan selection in BDI agents". In: *IJCAI International Joint Conference on Artificial Intelligence*. Vol. 0, 178–184. doi:10.24963/ijcai.2017/26.

J. W. Creswell and C. N. Poth. 2018 - 2018. *Qualitative inquiry and research design : choosing among five approaches*. eng. (4th edition. ed.). SAGE, Los Angeles. ISBN: 9781506330204.

N. Crook and J. Corneli. 2021. "The Anatomy of moral agency: A theological and neuroscience inspired model of virtue ethics". *Cognitive Computation and Systems*, 3, 2, 109–122. doi:10.1049/ccs2.12024.

A. Dafoe, Y. Bachrach, G. Hadfield, E. Horvitz, K. Larson, and T. Graepel. May 2021. "Cooperative AI: machines must learn to find common ground". en. *Nature*, 593, 7857, (May 2021), 33–36. Bandiera_abtest: a Cg_type: Comment Number: 7857 Publisher: Nature Publishing Group Subject_term: Machine learning, Computer science, Society, Technology, Sociology, Human behaviour. doi:10.1038/d41586-021-01170-0.

A. Davoust and M. Rovatsos. 2020. "Social contracts for non-cooperative games". In: *AIES 2020 - Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 43–49. doi:10.1145/3375627.3375829.

B. De Carolis, I. Mazzotta, and N. Novielli. 2010. *Enhancing conversational access to information through a socially intelligent agent*. Studies in Computational Intelligence. Vol. 301. Pages: 20. doi:10.1007/978-3-642-14000-6_1.

V. Dignum et al.. Dec. 2018. "Ethics by Design: Necessity or Curse?" en. In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, New Orleans LA USA, (Dec. 2018), 60–66. ISBN: 978-1-4503-6012-8. doi:10.1145/3278721.3278745.

P. Dognin et al.. 2025. "Contextual value alignment". In: *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

B. R. Duffy. 2003. "Anthropomorphism and the social robot". *Robotics and autonomous systems*, 42, 3-4, 177–190.

P. Eckersley. 2019. "Impossibility and uncertainty theorems in AI value alignment or why your AGI should not have a utility function". In: *CEUR Workshop Proceedings*. Vol. 2301. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85060611282&partnerID=40&md5=556a373ca14f2b53146d7d313ba556de.

A. K. Faulhaber, A. Dittmer, F. Blind, M. A. Wächter, S. Timm, L. R. Sütfeld, A. Stephan, G. Pipa, and P. König. Apr. 2019. "Human Decisions in Moral Dilemmas are Largely Described by Utilitarianism: Virtual Car Driving Study Provides Guidelines for Autonomous Driving Vehicles". en. *Science and Engineering Ethics*, 25, 2, (Apr. 2019), 399–418. doi:10.1007/s11948-018-0020-x.

M. Feffer, M. Skirpan, Z. Lipton, and H. Heidari. 2023. "From Preference Elicitation to Participatory ML: A Critical Survey & Guidelines for Future Research". In: *AIES 2023 - Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 38–48. doi:10.1145/3600211.3604661.

J. Fereday and E. Muir-Cochrane. 2006. "Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development". *International journal of qualitative methods*, 5, 1, 80–92.

S. G. Ficici, A. Pfeffer, and M.-D. Laboratory. 2008. "Simultaneously Modeling Humans' Preferences and their Beliefs about Others' Preferences". en.

E. Firt. 2023. "Calibrating machine behavior: a challenge for AI alignment". *Ethics and Information Technology*, 25, 3. doi:10.1007/s10676-023-09716-8.

J. Fisac et al.. 2020. "Pragmatic-Pedagogic Value Alignment". *Springer Proceedings in Advanced Robotics*, 10, 49–57. doi:10.1007/978-3-030-28619-4_7.

A. Franzke. 2022. "An exploratory qualitative analysis of AI ethics guidelines". *Journal of Information, Communication and Ethics in Society*, 20, 4, 401–423. doi:10.1108/JICES-12-2020-0125.

B. Friedman. 1996. "Value-sensitive design". *ACM Interactions*, 3, 6, 16–23. Retrieved June 30, 2020 from https://doi.org/10.1145/242485.242493.

I. Gabriel. 2020. "Artificial Intelligence, Values, and Alignment". *Minds and Machines*, 30, 3, 411–437. doi:10.1007/s11023-020-09539-2.

P. Gamez, D. Shank, C. Arnold, and M. North. 2020. "Artificial virtue: the machine question and perceptions of moral character in artificial moral agents". *AI and Society*, 35, 4, 795–809. doi:10.1007/s00146-020-00977-1.

C. Gavidia-Calderon, A. Bennaceur, A. Kordoni, M. Levine, and B. Nuseibeh. 2022. "What Do You Want From Me? Adapting Systems to the Uncertainty of Human Preferences". In: *Proceedings - International Conference on Software Engineering*, 126–130. doi:10.1109/ICSE-NIER55298.2022.9793539.

M. J. Gelfand, S. Gavrilets, and N. Nunn. 2024. "Norm dynamics: Interdisciplinary perspectives on social norm emergence, persistence, and change". *Annual Review of Psychology*, 75, 1, 341–378.

L. Gomez Tobon and E. Law. 2025. "Values in the loop: Designing interactive optimization with conversational feedback". In: *Proceedings of the 7th ACM Conference on Conversational User Interfaces*, 1–5.

N. Govindarajulu, R. Ghosh, S. Bringsjord, and V. Sarathy. 2019. "Toward the engineering of virtuous machines". In: *AIES 2019 - Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 29–35. doi:10.1145/3306618.3314256.

J. Graham, J. Haidt, S. Koleva, M. Motyl, R. Iyer, S. P. Wojcik, and P. H. Ditto. 2013. "Moral foundations theory: The pragmatic validity of moral pluralism". In: *Advances in experimental social psychology*. Vol. 47. Elsevier, 55–130.

J. Haas. 2020. "Moral Gridworlds: A Theoretical Proposal for Modeling Artificial Moral Cognition". *Minds and Machines*, 30, 2, 219–246. doi:10.1007/s11023-020-09524-9.

D. Hadfield-Menell, A. Dragan, P. Abbeel, and S. Russell. 2016. "Cooperative inverse reinforcement learning". In: *Advances in Neural Information Processing Systems*, 3916–3924. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85018873749&partnerID=40&md5=60d0969695bd71e2eeaa97a256706d97.

D. Hadfield-Menell and G. Hadfield. 2019. "Incomplete contracting and AI alignment". In: *AIES 2019 - Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 417–422. doi:10.1145/3306618.3314250.

T. Hagendorff. Mar. 2020. "The Ethics of AI Ethics: An Evaluation of Guidelines". en. *Minds and Machines*, 30, 1, (Mar. 2020), 99–120. doi:10.1007/s11023-020-09517-8.

S. Han, E. Kelly, S. Nikou, and E.-O. Svee. 2022. "Aligning artificial intelligence with human values: reflections from a phenomenological perspective". *AI and Society*, 37, 4, 1383–1395. doi:10.1007/s00146-021-01247-4.

H. Harland, R. Dazeley, B. Nakisa, F. Cruz, and P. Vamplew. 2023. "AI apology: interactive multi-objective reinforcement learning for human-aligned AI". *Neural Computing and Applications*, 35, 23, 16917–16930. doi:10.1007/s00521-023-08586-x.

T. Heyder, N. Passlack, and O. Posegga. 2023. "Ethical management of human-AI interaction: Theory development review". *Journal of Strategic Information Systems*, 32, 3. doi:10.1016/j.jsis.2023.101772.

A. Holgado-Sánchez. 2023. *Value-Awareness Engineering: Towards Learning Context-Based Value Taxonomies*. Vol. 14282 LNAI. Pages: 485. doi:10.1007/978-3-031-43264-4_35.

A. Holgado-Sánchez, J. Arias, M. Moreno-Rebato, and S. Ossowski. 2023. *On Admissible Behaviours for Goal-Oriented Decision-Making of Value-Aware Agents*. Vol. 14282 LNAI. Pages: 424. doi:10.1007/978-3-031-43264-4_27.

K. Huang, X. Liu, Q. Guo, T. Sun, J. Sun, Y. Wang, Z. Zhou, Y. Wang, Y. Teng, X. Qiu, et al.. 2023. "Flames: Benchmarking value alignment of llms in chinese". *arXiv preprint arXiv:2311.06899*.

K. Huang, X. Liu, Q. Guo, T. Sun, J. Sun, Y. Wang, Z. Zhou, Y. Wang, Y. Teng, X. Qiu, et al.. June 2024. "Flames: Benchmarking Value Alignment of LLMs in Chinese". In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Ed. by K. Duh, H. Gomez, and S. Bethard. Association for Computational Linguistics, Mexico City, Mexico, (June 2024), 4551–4591. doi:10.18653/v1/2024.naacl-long.256.

S. Jasanoff. 2016. *The ethics of invention: Technology and the human future*. WW Norton & Company.

M. Juric, A. Sandic, and M. Brcic. 2020. "AI safety: State of the field through quantitative lens". In: *2020 43rd International Convention on Information, Communication and Electronic Technology, MIPRO 2020 - Proceedings*, 1254–1259. doi:10.23919/MIPRO48935.2020.9245153.

A. A. Khan, S. Badshah, P. Liang, M. Waseem, B. Khan, A. Ahmad, M. Fahmideh, M. Niazi, and M. A. Akbar. June 2022. "Ethics of AI: A Systematic Literature Review of Principles and Challenges". In: *Proceedings of the 26th International Conference on Evaluation and Assessment in Software Engineering* (EASE '22). Association for Computing Machinery, New York, NY, USA, (June 2022), 383–392. ISBN: 978-1-4503-9613-4. doi:10.1145/3530019.3531329.

A. Kierans, A. Ghosh, H. Hazan, and S. Dori-Hacohen. 2025. "Quantifying misalignment between agents: Towards a sociotechnical understanding of alignment". In: *Proceedings of the AAAI Conference on Artificial Intelligence*, 27365–27373.

O. J. Kwon, D. E. Matsunaga, and K.-E. Kim. 2024. "GDPO: Learning to Directly Align Language Models with Diversity Using GFlowNets". *arXiv preprint arXiv:2410.15096*.

B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman. 2017. "Building machines that learn and think like people". *Behavioral and brain sciences*, 40, e253.

R. Lera-Leri, F. Bistaffa, M. Serramia, M. Lopez-Sanchez, and J. Rodriguez-Aguilar. 2022. "Towards Pluralistic Value Alignment: Aggregating Value Systems through ℓ_p-Regression". In: *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS*. Vol. 2, 780–788. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85134326448&partnerID=40&md5=3dd59846082f36be0280379d03447d35.

O. Li. 2021. "Problems with "Friendly AI"". *Ethics and Information Technology*, 23, 3, 543–550. doi:10.1007/s10676-021-09595-x.

E. Liscio, R. Lera-Leri, F. Bistaffa, R. Dobbe, C. Jonker, M. Lopez-Sanchez, J. Rodriguez-Aguilar, and P. Murukannaiah. 2023. "Value Inference in Sociotechnical Systems". In: *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS*. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85171269701&partnerID=40&md5=5391a2f951325b423b18c4afbcc103bc.

E. Liscio. 2021. "A Collaborative Platform for Identifying Context-Specific Values". en.

E. Liscio, M. van der Meer, L. C. Siebert, C. M. Jonker, and P. K. Murukannaiah. Mar. 2022. "What values should an agent align with?" en. *Autonomous Agents and Multi-Agent Systems*, 36, 1, (Mar. 2022), 23. doi:10.1007/s10458-022-09550-0.

A. Loreggia, N. Mattei, F. Rossi, and K. Venable. 2019. "Metric learning for value alignment". In: *CEUR Workshop Proceedings*. Vol. 2419. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85071011100&partnerID=40&md5=9c1e49f47cd12db5db9008ac44a128d8.

S. Maitra. 2020. "Artificial intelligence and indigenous perspectives: Protecting and empowering intelligent human beings". In: *AIES 2020 - Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 320–326. doi:10.1145/3375627.3375845.

J. Martınez-Miranda and A. Aldea. 2005. "Emotions in human and artificial intelligence". *Computers in human behavior*, 21, 2, 323–341.

E. Mason. 2006. "Value pluralism".

M. Mechergui and S. Sreedharan. 2023. "Goal Alignment: Re-analyzing Value Alignment Problems Using Human-Aware AI". In: *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS*. Vol. 2023-May, 2331–2333. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85171284300&partnerID=40&md5=0148f2096ca35887f629eb14b5d65e78.

C. Meske, E. Bunde, J. Schneider, and M. Gersch. 2022. "Explainable artificial intelligence: objectives, stakeholders, and future research opportunities". *Information Systems Management*, 39, 1, 53–63.

S. Milli, D. Hadfield-Menell, A. Dragan, and S. Russell. 2017. "Should robots be obedient?" In: *IJCAI International Joint Conference on Artificial Intelligence*. Vol. 0, 4754–4760. doi:10.24963/ijcai.2017/662.

N. Montes, N. Osman, C. Sierra, and M. Slavkovik. Feb. 2023. *Value Engineering for Autonomous Agents*. arXiv:2302.08759 [cs]. (Feb. 2023). doi:10.48550/arXiv.2302.08759.

A. Morris-Martin, M. De Vos, and J. Padget. 2019. "Norm emergence in multiagent systems: a viewpoint paper". *Autonomous Agents and Multi-Agent Systems*, 33, 706–749.

M. J. Muller and S. Kuhn. 1993. "Participatory design". *Communications of the ACM*, 36, 6, 24–28.

G. Murray. 2017. *Stoic ethics for artificial agents*. Vol. 10233 LNAI. Pages: 384. doi:10.1007/978-3-319-57351-9_42.

M. Naeem, W. Ozuem, K. Howell, and S. Ranfagni. 2023. "A step-by-step process of thematic analysis to develop a conceptual model in qualitative research". *International Journal of Qualitative Methods*, 22, 16094069231205789.

S. Nikolaidis, Y. Zhu, D. Hsu, and S. Srinivasa. 2017. "Human-Robot Mutual Adaptation in Shared Autonomy". In: *ACM/IEEE International Conference on Human-Robot Interaction*. Vol. Part F127194, 294–302. doi:10.1145/2909824.3020252.

P. Noriega, H. Verhagen, J. Padget, and M. d'Inverno. 2022. *Design Heuristics for Ethical Online Institutions*. Vol. 13549 LNAI. Pages: 230. doi:10.1007/978-3-031-20845-4_14.

P. Noriega and E. Plaza. 2022. "The Use of Agent-based Simulation of Public Policy Design to Study the Value Alignment Problem". en.

P. Noriega, H. Verhagen, J. Padget, and M. d'Inverno. 2023. "Addressing the Value Alignment Problem Through Online Institutions". en. In: *Coordination, Organizations, Institutions, Norms, and Ethics for Governance of Multi-Agent Systems XVI* (Lecture Notes in Computer Science). Ed. by N. Fornara, J. Cheriyan, and A. Mertzani. Springer Nature Switzerland, Cham, 77–94. ISBN: 978-3-031-49133-7. doi:10.1007/978-3-031-49133-7_5.

N. Osman and M. d'Inverno. 2024. "A computational framework of human values".

O. Osoba, B. Boudreaux, and D. Yeung. 2020. "Steps towards value-aligned systems". In: *AIES 2020 - Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 332–336. doi:10.1145/3375627.3375872.

I. Padhi, K. N. Ramamurthy, P. Sattigeri, M. Nagireddy, P. Dognin, and K. R. Varshney. 2024. "Value alignment from unstructured text". *arXiv preprint arXiv:2408.10392*.

U. Pagallo. 2016. *Even angels need the rules: AI, roboethics, and the law*. Vol. 285. Pages: 215. doi:10.3233/978-1-61499-672-9-209.

M. J. Page et al.. 2021. "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews". *bmj*, 372.

M. Peschl, A. Zgonnikov, F. Oliehoek, and L. Siebert. 2022. "MORAL: Aligning AI with Human Norms through Multi-Objective Reinforced Active Learning". In: *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS*. Vol. 2, 1038–1046. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85134357119&partnerID=40&md5=0c2fe9165d7d23bdf1c8dabf1a87665a.

M. Peterson. 2019. "The value alignment problem: a geometric approach". *Ethics and Information Technology*, 21, 1, 19–28. doi:10.1007/s10676-018-9486-0.

K. Pigmans, H. Aldewereld, V. Dignum, and N. Doorn. 2017. *The role of values*. Vol. 10315 LNAI. Pages: 148. doi:10.1007/978-3-319-66595-5_8.

I. van de Poel. Sept. 2020. "Embedding Values in Artificial Intelligence (AI) Systems". en. *Minds and Machines*, 30, 3, (Sept. 2020), 385–409. doi:10.1007/s11023-020-09537-4.

I. Rahwan. 2018. "Society-in-the-loop: programming the algorithmic social contract". *Ethics and Information Technology*, 20, 1, 5–14. doi:10.1007/s10676-017-9430-8.

A. Rapoport and A. M. Chammah. 1965. *Prisoner's dilemma: A study in conflict and cooperation*. Vol. 165. University of Michigan press.

W. Ratoff. 2021. "Can the predictive processing model of the mind ameliorate the value-alignment problem?" *Ethics and Information Technology*, 23, 4, 739–750. doi:10.1007/s10676-021-09611-0.

G. Reichberg and H. Syse. 2021. "Applying AI on the battlefield: The ethical debates". In: *Robotics, AI, and Humanity: Science, Ethics, and Policy*, 147–159. doi:10.1007/978-3-030-54173-6_12.

P. Robinson. 2023. "Action Guidance and AI Alignment". In: *AIES 2023 - Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 387–395. doi:10.1145/3600211.3604714.

M. Rodriguez-Soto, M. Lopez-Sanchez, and J. Rodriguez-Aguilar. 2021. "Guaranteeing the Learning of Ethical Behaviour through Multi-Objective Reinforcement Learning". In: *ALA 2021 - Adaptive and Learning Agents Workshop at AAMAS 2021*. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85164790058&partnerID=40&md5=8c6e41a552a28873676f41c65dd8937a.

M. Rodriguez-Soto, M. Serramia, M. Lopez-Sanchez, and J. Rodriguez-Aguilar. 2022. "Instilling moral value alignment by means of multi-objective reinforcement learning". *Ethics and Information Technology*, 24, 1. doi:10.1007/s10676-022-09635-0.

M. Rokeach. 1967. "Rokeach value survey". *The nature of human values*.

S. Rosenthal and M. Veloso. 2012. "Monte Carlo preference elicitation for learning additive reward functions". In: *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication*, 886–891. doi:10.1109/ROMAN.2012.6343863.

J.-J. Rousseau. 2016. "The social contract". In: *Democracy: A Reader*. Columbia University Press, 43–51.

S. Russell. 2019. *Human compatible: AI and the problem of control*. Penguin Uk.

S. Russell, D. Dewey, and M. Tegmark. Dec. 2015. "Research Priorities for Robust and Beneficial Artificial Intelligence". en. *AI Magazine*, 36, 4, (Dec. 2015), 105–114. Number: 4. doi:10.1609/aimag.v36i4.2577.

A. Salles, K. Evers, and M. Farisco. 2020. "Anthropomorphism in AI". *AJOB neuroscience*, 11, 2, 88–95.

S. A. Salloum, K. M. Alomari, A. M. Alfaisal, R. A. Aljanada, and A. Basiouni. 2025. "Emotion recognition for enhanced learning: using AI to detect students' emotions and adjust teaching methods". *Smart Learning Environments*, 12, 1, 21.

L. Sanneman and J. Shah. 2023a. "Transparent Value Alignment". In: *ACM/IEEE International Conference on Human-Robot Interaction*, 557–560. doi:10.1145/3568294.3580147.

L. Sanneman and J. Shah. 2023b. "Validating metrics for reward alignment in human-autonomy teaming". *Computers in Human Behavior*, 146. doi:10.1016/j.chb.2023.107809.

G. Sarma, N. Hay, and A. Safron. 2018. *AI safety and reproducibility: Establishing robust foundations for the neuropsychology of human values*. Vol. 11094 LNCS. Pages: 512. doi:10.1007/978-3-319-99229-7_45.

M. Scheutz. 2011. "13 The inherent dangers of unidirectional emotional bonds between humans and social robots". *Robot ethics: The ethical and social implications of robotics*, 205.

S. H. Schwartz. 2012. "An overview of the Schwartz theory of basic values". *Online readings in Psychology and Culture*, 2, 1, 11.

S. H. Schwartz. 1992. "Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries". In: *Advances in experimental social psychology*. Vol. 25. Elsevier, 1–65.

S. H. Schwartz and W. Bilsky. 1987. "Toward a universal psychological structure of human values." *Journal of personality and social psychology*, 53, 3, 550.

A. Sen. N.d. "Amartya Sen on well-being, critical voice and social choice theory".

M. Serramia, M. Lopez-Sanchez, and J. Rodriguez-Aguilar. 2020. "A qualitative approach to composing value-aligned norm systems". In: *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS*. Vol. 2020-May, 1233–1241. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85096672634&partnerID=40&md5=b92d419125e110b4b4a102fbcaca1c31.

M. Serramia, M. Rodriguez-Soto, M. Lopez-Sanchez, J. A. Rodriguez-Aguilar, F. Bistaffa, P. Boddington, M. Wooldridge, and C. Ansotegui. 2023. "Encoding ethics to compute value-aligned norms". *Minds and Machines*, 33, 4, 761–790.

N. P. Shaw, A. Stöckel, R. W. Orr, T. F. Lidbetter, and R. Cohen. Dec. 2018. "Towards Provably Moral AI Agents in Bottom-up Learning Frameworks". In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (AIES '18). Association for Computing Machinery, New York, NY, USA, (Dec. 2018), 271–277. ISBN: 978-1-4503-6012-8. doi:10.1145/3278721.3278728.

L. C. Siebert, E. Liscio, P. K. Murukannaiah, L. Kaptein, S. Spruit, J. van den Hoven, and C. Jonker. 2022. "Estimating Value Preferences in a Hybrid Participatory System". In: *HHAI2022: Augmenting Human Intellect*. IOS Press, 114–127. doi:10.3233/FAIA220193.

C. Sierra, N. Osman, P. Noriega, J. Sabater-Mir, and A. Perelló. Oct. 2021. *Value alignment: a formal approach*. arXiv:2110.09240 [cs]. (Oct. 2021). doi:10.48550/arXiv.2110.09240.

A. Sison and D. Redín. 2023. "A neo-aristotelian perspective on the need for artificial moral agents (AMAs)". *AI and Society*, 38, 1, 47–65. doi:10.1007/s00146-021-01283-0.

J. J. Smith, A. Buhayh, A. Kathait, P. Ragothaman, N. Mattei, R. Burke, and A. Voida. 2023. "The Many Faces of Fairness: Exploring the Institutional Logics of Multistakeholder Microlending Recommendation". In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '23). Association for Computing Machinery, Chicago, IL, USA, 1652–1663. ISBN: 9798400701924. doi:10.1145/3593013.3594106.

N. Soares. July 2018. "The Value Learning Problem". en. In: *Artificial Intelligence Safety and Security*. (1st ed.). Ed. by R. V. Yampolskiy. Chapman and Hall/CRC, First edition. | Boca Raton, FL : CRC Press/Taylor & Francis Group, 2018., (July 2018), 89–97. ISBN: 978-1-351-25138-9. doi:10.1201/9781351251389-7.

N. Soares, B. Fallenstein, S. Armstrong, and E. Yudkowsky. 2015. "Corrigibility". In: *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*.

M. S. Spektor, S. Bhatia, and S. Gluth. 2021. "The elusiveness of context effects in decision making". *Trends in Cognitive Sciences*, 25, 10, 843–854.

J. Stenseke. 2023. "Artificial virtuous agents: from theory to machine implementation". *AI and Society*, 38, 4, 1301–1320. doi:10.1007/s00146-021-01325-7.

M. Sutrop. Dec. 2020. "Challenges of Aligning Artificial Intelligence with Human Values". en. *Acta Baltica Historiae et Philosophiae Scientiarum*, 8, 2, (Dec. 2020), 54–72. doi:10.11590/abhps.2020.2.04.

J. Svegliato, S. Nashed, and S. Zilberstein. 2020. "Ethically compliant planning in moral autonomous systems". In: *CEUR Workshop Proceedings*. Vol. 2640. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85089626296&partnerID=40&md5=b3d6a75fdc806263aae90b6e0348ad9f.

J. Szabo, J. Such, and N. Criado. 2020. *Understanding the Role of Values and Norms in Practical Reasoning*. Vol. 12520 LNAI. Pages: 439. doi:10.1007/978-3-030-66412-1_27.

J. Szabo, J. Such, N. Criado, and S. Modgil. 2022. *Integrating Quantitative and Qualitative Reasoning for Value Alignment*. Vol. 13442 LNAI. Pages: 402. doi:10.1007/978-3-031-20614-6_22.

A. Theodorou and V. Dignum. Jan. 2020. "Towards ethical and socio-legal governance in AI". *Nature Machine Intelligence*, 2, 1, (Jan. 2020), 10–12. doi:10.1038/s42256-019-0136-y.

J. J. Thomson. 1984. "The trolley problem". *Yale LJ*, 94, 1395.

S. Thornton, S. Pan, S. Erlien, and J. Gerdes. 2017. "Incorporating Ethical Considerations into Automated Vehicle Control". *IEEE Transactions on Intelligent Transportation Systems*, 18, 6, 1429–1439. doi:10.1109/TITS.2016.2609339.

S. Tolmeijer, M. Kneer, C. Sarasua, M. Christen, and A. Bernstein. Dec. 2021. "Implementations in Machine Ethics: A Survey". *ACM Computing Surveys*, 53, 6, (Dec. 2021), 132:1–132:38. doi:10.1145/3419633.

R. Tomsett, D. Braines, D. Harborne, A. Preece, and S. Chakraborty. 2018. "Interpretable to whom? A role-based model for analyzing interpretable machine learning systems". *arXiv preprint arXiv:1806.07552*.

S.-T. Tzeng. 2022. "Engineering Normative and Cognitive Agents with Emotions and Values". en.

S. Umbrello and I. van de Poel. Aug. 2021. "Mapping value sensitive design onto AI for social good principles". en. *AI and Ethics*, 1, 3, (Aug. 2021), 283–296. doi:10.1007/s43681-021-00038-3.

S. Vallor. Mar. 2015. "Moral Deskilling and Upskilling in a New Machine Age: Reflections on the Ambiguous Future of Character". en. *Philosophy & Technology*, 28, 1, (Mar. 2015), 107–124. doi:10.1007/s13347-014-0156-9.

P. Vamplew, R. Dazeley, C. Foale, S. Firmin, and J. Mummery. 2018. "Human-aligned artificial intelligence is a multiobjective problem". *Ethics and Information Technology*, 20, 1, 27–40. doi:10.1007/s10676-017-9440-6.

I. Van de Poel. 2013. "Translating values into design requirements". *Philosophy and engineering: Reflections on practice, principles and process*, 253–266.

W. Visser, K. Hindriks, and C. Jonker. 2011. "Interest-based preference reasoning". In: *ICAART 2011 - Proceedings of the 3rd International Conference on Agents and Artificial Intelligence*. Vol. 1, 79–88. https://www.scopus.com/inward/record.uri?eid=2-s2.0-79960115725&partnerID=40&md5=f0ae5dc06e0cc017b1a603866e618fea.

W. Wallach and C. Allen. Nov. 2008. *Moral Machines: Teaching Robots Right from Wrong*. en. Google-Books-ID: _r3N82ETng4C. Oxford University Press, (Nov. 2008). ISBN: 978-0-19-970596-2.

M. Waser. 2015. "Designing, Implementing and Enforcing a Coherent System of Laws, Ethics and Morals for Intelligent Machines (Including Humans)". In: *Procedia Computer Science*. Vol. 71, 106–111. doi:10.1016/j.procs.2015.12.213.

M. Waser and D. Kelley. 2018. "Architecting a human-like emotion-driven conscious moral mind for value alignment and AGI safety". In: *AAAI Spring Symposium - Technical Report*. Vol. 2018-March, 88–94. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85102583382&partnerID=40&md5=cd98ce5035085a4a28e1f931c70b17d3.

S. Xu, W. Dong, Z. Guo, X. Wu, and D. Xiong. 2024. "Exploring Multilingual Human Value Concepts in Large Language Models: Is Value Alignment Consistent, Transferable and Controllable across Languages?" *CoRR*, abs/2402.18120. https://doi.org/10.48550/arXiv.2402.18120.

H. Ye, Y. Xie, Y. Ren, H. Fang, X. Zhang, and G. Song. Apr. 2025. "Measuring human and AI values based on generative psychometrics with large language models". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 39. (Apr. 2025), 26400–26408.

E. Yudkowsky. 2016. "The AI Alignment Problem: Why It's Hard, and Where to Start". en.

S. Zhuang and D. Hadfield-Menell. 2020. "Consequences of Misaligned AI". In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 15763–15773. Retrieved Dec. 15, 2023 from https://proceedings.neurips.cc/paper/2020/hash/b607ba543ad05417b8507ee86c54fcb7-Abstract.html.

L. M. Zintgraf, D. M. Roijers, S. Linders, C. M. Jonker, and A. Nowé. Feb. 2018. *Ordered Preference Elicitation Strategies for Supporting Multi-Objective Decision Making*. arXiv:1802.07606 [cs, stat]. (Feb. 2018). doi:10.48550/arXiv.1802.07606.

J. Zoshak and K. Dew. 2021. "Beyond kant and bentham: How ethical theories are being used in artificial moral agents". In: *Conference on Human Factors in Computing Systems - Proceedings*. doi:10.1145/3411764.3445102.

T. Zurek, M. Araszkiewicz, and D. Stachura-Zurek. 2022. "Reasoning with principles". *Expert Systems with Applications*, 210. doi:10.1016/j.eswa.2022.118496.

T. Zurek and M. Mokkas. 2021. "Value-based reasoning in autonomous agents". *International Journal of Computational Intelligence Systems*, 14, 1, 896–921. doi:10.2991/IJCIS.D.210203.001.

A Search Details

Our database of papers from *Scopus* was initially constructed from three searches. We selected papers to include in the review by first by examining the title and abstract, and then by examining the full text. Further papers were obtained by examining the bibliographies of papers included in the first coding pass.

A.1 Initial Value Alignment Search

Search Terms:

TITLE-ABS-KEY ("Value Align*" OR "Alignment Problem*" OR "Value-based reasoning" OR "Human Preference*") AND TITLE-ABS-KEY ("Artificial Intelligence*" OR "AI*" OR "Agent" OR "Autonomous" OR "Intelligent System") AND NOT ("time-series" OR "knowledge-graph" OR "SINS" OR "protein sequence*" OR "genome*") AND

PUBYEAR > 1900 AND PUBYEAR < 2024 AND (LIMIT-TO (SUBJAREA , "COMP")) AND (LIMIT-TO (PUBSTAGE , "final")) AND (LIMIT-TO (LANGUAGE , "English"))

Date of Search: November 7th 2023
Initial Results: 535
Included results after abstract & title check: 162
Included results after full text check: 90

Our first set of TITLE-ABS-KEY terms aimed to identify value alignment related papers, or those that connect to human preferences. Our second set of TITLE-ABS-KEY terms restricts this to papers discussing artificial intelligence related systems. Our excluding TITLE-ABS-KEY terms are based on a previous scoping review we conducted. We used them to exclude irrelevant topics that trigger based on our searching for alignment in computer science. We then filter by publication year and subject area, and the last two criteria restrict our search to papers that have been quality assessed through being at final publication stage and will be in a language we can work with.

A.2 Virtue Ethics and Social Contracts Search

Search Terms:
TITLE-ABS-KEY ("Virtue ethic*" OR "Social contract") AND TITLE-ABS-KEY ("Artificial Intelligence*" OR "AI*" OR "Agent" OR "Autonomous" OR "Intelligent System") AND NOT ("time-series" OR "knowledge-graph" OR "SINS" OR "protein sequence*" OR "genome*") AND PUBYEAR > 1900 AND PUBYEAR < 2024 AND (LIMIT-TO (SUBJAREA , "COMP")) AND (LIMIT-TO (PUBSTAGE , "final")) AND (LIMIT-TO (LANGUAGE , "English"))

Date of Search: November 9th 2023
Initial Results: 138
Included results after abstract & title check: 46
Included results after full text check: 31

Our second search replaced the first set of TITLE-ABS-KEY terms in order to ensure coverage of virtue ethics and the social contract, as these were deemed relevant to the value alignment problem. The rest of the search followed the same reasoning as the first search.

A.3 Multi-Agent Systems Search

Search Terms:
TITLE-ABS-KEY ("Value Align*" OR "Alignment Problem*" OR "Value-based reasoning" OR "Human Preference*") AND TITLE-ABS-KEY ("Multi Agent System" OR "Agent Based") AND NOT ("time-series" OR "knowledge-graph" OR "SINS" OR "protein sequence*" OR "genome*") AND PUBYEAR > 1900 AND PUBYEAR < 2024 AND (LIMIT-TO (SUBJAREA , "COMP")) AND (LIMIT-TO (PUBSTAGE , "final")) AND (LIMIT-TO (LANGUAGE , "English"))

Date of Search: February 2nd 2024
Initial Results: 53

This third search was performed after the initial phases of coding indicated a strong motivation for considering the value alignment problem through a multi-agent system perspective. The purpose of this search was to ensure that we properly integrated thought from multi-agent systems into the review, and to avoid replicating existing work. This was again achieved by modifying the second set of TITLE-ABS-KEY terms, while keeping the rest of the search the same.

We only included 8 papers from this search, as the rest of the papers had either already been added from the previous two searches or were irrelevant.

B Inclusion/Exclusion Criteria

B.1 Abstract & Title Check Criteria

Ex1_Moral Values and Preferences

Is the paper discussing values in the context of moral and social values, rather than only technical values (e.g. measurements and function outputs)?

We defined moral/social values as factors that lead to preferences in humans or otherwise help coordinate group decisions.

We did not require an attempt to define or measure these values to have been made in the paper.

The purpose of this criteria was to exclude papers that only used the term value in the sense of metrics, without looking at the type of values we were interested in for this survey.

Ex2_Values and Preferences Learning

Is the paper discussing the challenge of integrating values, feasibly in the form of preferences, in forms of technology?

The purpose of this criteria was to include papers that linked values to technology. Our inclusion of preferences as a viable expression of values was based on our prior knowledge before starting the survey.

B.2 Full Text Check Criteria

In1_Accessibility

Do we have access to a copy of the paper?

This included open-source access or institutional access. It did not include access that could be obtained through a separate purchase or similar mechanic.

In2_Value_Alignment

Does the paper discuss the challenge of getting humans and autonomous systems to act with aligned behaviour?

We defined aligned behaviour here as behaviour that is considered acceptable by humans according to their moral/social values.

At this stage of the research, we did not discriminate between alignment being targeted at a single human or multiple humans.

We allowed discussions of multiple agents without specifying which agents were human and which were autonomous, as long as there was no reason this could not include mixtures of autonomous agents and humans.

We excluded alignment discussed only in the context of physically coordinated actions, as was common in human-robot interaction papers.

In3_Definitions

Does the paper include a definition of value alignment to some extent, even if it does not explicitly call it such?

This referred to an attempt to explain the meaning of the value alignment process, as per our working understanding of it (humans and AI agents acting in ways that agree with each other), beyond simply indicating that it is a problem that exists.

We included indirect definition through proposing a methodology for solving the problem. We interpreted this as the proposed solution through this methodology indicating the criteria for solving value alignment, in whole or partially.

B.3 Full Coding Pass Criteria

Ex4_Value Aligned AI Governance

We did not include topics focusing on the governance around value aligned AI. This was outside the scope of this survey.

Ex5_Moral Status of AI

We did not include papers focusing on the moral status or moral capabilities of AI. This was outside the scope of this survey.

Incl6_Implementation method OR Design Discussion OR VA Modelling

Does the paper contain a theoretical/practical implementation of value alignment?

OR

Does the paper contain discussion around the design of value aligned systems, including overall structure or individual components?

OR

Does this paper contain a discussed or implemented model of value aligned systems?

The intent of this criteria was to select papers that included practical components of value alignment or contributed discussion to how these components should be designed.

Ex7_Specific Values

We did not include papers that only focused on specific values, such as fairness or explainability. Our goal was to understand the process generally, rather than in the context of a specific value.

In8_Normative Behaviour

We included papers that discussed social, normative or some other organisational alignment of agents. This was done to include papers focusing on normative modelling, as we considered them relevant to understanding the value alignment process.

In_Final

Will we include the article in the review?

We fully coded the paper if it passed all exclusion checks in the full coding pass criteria checks, and at least one of the inclusion checks in the full coding pass criteria.

Received 10 April 2025; accepted 20 November 2025