Ethical Alignment in Citizen-Centric AI

Jayati Deshmukh[0000−0002−1144−2635], Vahid Yazdanpanah[0000−0002−4468−6193], Sebastian Stein[0000−0003−2858−8857], and Timothy Norman[0000−0002−6387−4034]

University of Southampton, Southampton, UK
{J.Deshmukh,V.Yazdanpanah,S.Stein,T.J.Norman}@soton.ac.uk

Abstract. This paper discusses the importance of ethical alignment in AI systems, particularly those designed with citizen end users in mind. It explores the intersection of responsible AI, socio-technical systems, and citizen-centric design, proposing that addressing the *ethical* aspect of decisions in citizen-centric AI systems enhances trust and acceptance of AI technologies. We focus on four key areas: (1) the formal specification of ethical principles, (2) processes to extract and elicit individual users' ethical preferences, (3) aggregating these ethical preferences for a collective, and (4) mechanisms to ensure that the behaviour of AI systems aligns with the collective ethical preferences. We put forward a research roadmap by identifying challenges in these areas and highlighting solution concepts with the potential to address them.

Keywords: ethical agents · multiagent systems · collectives · citizen-centric AI · trustworthy autonomous systems

1 Introduction

Ethics broadly define the key governing principles based on which we act. While ethics has been studied and discussed for thousands of years, the field of Artificial Intelligence (AI) is relatively new, being a few decades old. However, AI is getting deployed in many real-world applications across diverse sectors such as healthcare, transportation, agriculture and supply chain management [42,41]. As AI is advancing and impacting a large number of citizens and communities with diverse values, it is crucial to study the intersection of ethics and AI to ensure that these systems not only solve the technical problems, but also that it is ethically aligned and act responsibly while interacting with humans, other AI systems with different values, and the socio-technical fabric of society.

To ensure responsible and ethical behaviour of AI systems, *responsible AI* is an active research area that focuses on building AI systems that are designed to act in alignment with societal norms and values [21]. Moreover, there exists a rich body of work on socio-technical systems focusing on the interplay between technology and society as well as means to govern the behaviour of such systems and ensure their alignment with societal values [54,30]. To complement these approaches, citizen-centric AI and multi-agent systems is a new research direction which keeps the citizens (as end users with limited technical skills) in

the centre of attention during the design, development and deployment of AI systems [50]. Citizen-centric systems should be aware of the diverse preferences[1] of users; sensitive to changing preferences and feedback received from users; focused on benefiting users (aligned with Russel's recent view on AI research [44]); and auditable and able to provide answers and explanations for their behaviour. All these research directions are important and are currently being pursued independently. We believe that ensuring ethical alignment of AI demands for more intersection between responsible and ethical AI (moral philosophy), socio-technical systems (social computing) and citizen-centric AI (computer science) and envision such a multidisciplinary research agenda to play a key role in the wider adoption of AI and successful deployment of ethical AI-based systems in real-world applications. In this paper, we present our research roadmap and put forward components key for designing AI-based systems that behave ethically and are aligned to the needs of citizen end users.

Studies have highlighted that AI systems which act ethically and are governed ethically have greater trust and acceptance in the citizens who use these AI systems [58]. Moreover, in our recent discussions with various experts and industry stakeholders developing or working with AI systems [1,2], a common point emerged regarding building AI-based systems that demonstrate responsible and ethical behaviour in the context in which they operate. Also, it should be ensured that the stakeholders who will use these systems are involved during development, so that they can trust these systems [49].

Throughout this paper, we consider various examples of AI-based systems, specifically where the ethical aspects are prominent in the citizen-centric systems. For example, we discuss challenges related to multiple stakeholders with different preferences and different models of ethics, sharing green energy or riding an autonomous vehicle, using different types of autonomous traffic lights which manage the traffic flows in a region and impact the driving experience of the drivers. Across these diverse scenarios, we highlight some of the existing challenges and different ways in which developing ethical citizen-centric systems might be able to address these challenges.

The overall workflow of our vision of an ethically aligned citizen-centric AI system is presented as a block diagram in Figure 1. Various paradigms of ethics are computationally represented using *Contextual Ethics Representation Methods*. Agents in different colours represent people with different paradigms of ethics. The *Ethics Elicitation Component* infers the ethical preferences of the involved citizens and maps them to the computational models of ethics. Next, the preferences of all the involved stakeholders are collated and combined using the *Ethics Aggregation Component*. Finally, the *Ethics Alignment Mechanism* validates the output of an AI-based system (like an autonomous vehicle, a smart energy meter etc.) to be ethically aligned with the aggregated ethical preferences of the stakeholders. Thus, our proposed approach of an ethically aligned citizen-

[1] Note that by "preference", we are referring to users' preference with respect to different ethical theories, e.g., if a user prefers their autonomous vehicle to follow the utilitarian view, then it is desirable for the vehicle to do so.

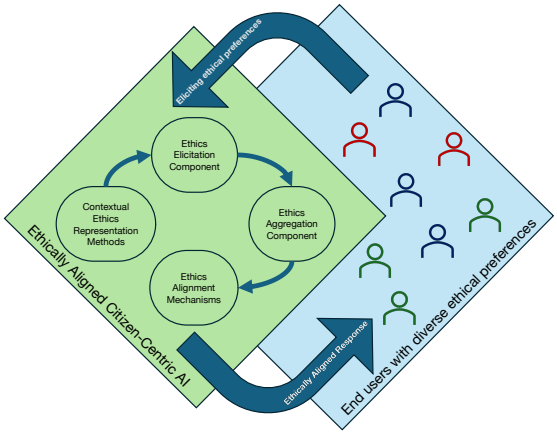

Fig. 1. Ethically Aligned Citizen-Centric AI: A context-specific AI system which interacts with users having diverse ethical preferences (represented using different colours).

centric system acts as a layer on top of any existing AI-based system to ensure that it acts ethically, aligned with everyone's ethical preferences.

The paper is organised as follows: In Section 2, different ethical paradigms which can be modelled in different AI applications are presented and Section 3 describes computational approaches to represent these paradigms of ethics in AI systems. Section 4 elaborates on different approaches to estimating and eliciting the ethical preferences of users. We look at different ways to aggregate the ethical preferences of a group or collective of agents in a system in Section 5. An AI-based system should be aligned with the underlying ethical preferences of its users; some of the key challenges to achieving it are discussed in Section 6. Finally, we bring it together as a case discussion in Section 7 and in Section 8, we conclude and present future directions on building ethical citizen-centric AI.

2 Ethical Paradigms: One Size Won't Fit All

There are various paradigms of ethics which have been studied and analysed over time. The most well-known approach is *normative ethics* [48], which looks at how one *ought to* act. From a moral point of view, it judges whether an action is right or wrong. There are primarily three theories of normative ethics: *utilitarianism*, which estimates the collective utility of actions and selects the one which gives the maximum utility to all; *virtue ethics*, which involves demonstrating some virtues which are relevant to the context; and *deontology*, which denotes fulfilling one's duties and following the norms. We will elaborate on how each of these is relevant in the context of citizen-centric AI and some of the existing challenges of incorporating ethics[2].

[2] In this work, we are not evaluating different paradigms of ethics, rather we aim to allow users to see their choice of ethics implemented in the AI technologies they use.

Utilitarianism looks at the consequences of actions on agents and the system as a whole [36,35,5]. An action is deemed ethical using this approach if it maximises the collective utility. All the stakeholders in a system are equally important and the utility of all agents is accounted for with equal weight. Different approaches look at the utility at different timescales from short-term to long-term and this results in variations across models. In this approach, it is difficult to estimate the utility of all the stakeholders in a system. Also, in most open-world systems, it is hard to know the consequences of actions in advance in order to estimate the total utility.

Virtue ethics [53] focuses on the agent itself rather than the intent behind or consequences of its actions. An agent is called ethical if it demonstrates the virtues relevant to the context in which it operates. For example, a soldier might demonstrate virtues like courage, bravery, and valour, and a healthcare worker might demonstrate virtues like care and empathy. Also, Aristotle elaborated that virtues should be demonstrated in moderation, in the *right* amount around the *golden mean* representing the middle ground between the extremes. Thus, in this perspective, agents should constantly adjust and ensure that they demonstrate all the required virtues to the optimum extent in a specific context. Virtue ethics also has some challenges– virtues are abstract, so it is difficult to measure virtues demonstrated by agents. It requires continuous effort from agents to demonstrate virtuous behaviour, and agents should be able to resolve conflicting virtues.

Deontological ethics emphasises following the rules and norms by agents. As long as the agent fulfils its obligations by following the rules, it is considered ethical, irrespective of the consequences of its actions. Immanuel Kant [31] specified two imperatives to design norms: *categorical imperative*: which specifies that a norm should be such that it can be a universal law, i.e. it leads to a better system state when everyone follows that norm; and *practical imperative*: which emphasises that people cannot be used as a means to achieve an end. Some of the challenges of this approach are: it is difficult to formulate the rules, especially in contexts where all possible outcomes are not known in advance; agents should be able to handle contradicting rules; sometimes agents need time to deliberate on the rules, which might not be possible when they need to make quick decisions.

For example, a smart home energy management system manages how, when and which energy to consume in a home, based on its user's preferences. Such a system aligning with users with different paradigms of ethics will operate differently. A *utilitarian energy agent* looks at not just its utility but the utility of the collective, which in this case might be the building. It should be able to operate such that the energy consumption of the building as a whole is optimised. A *virtue ethical energy agent* might be modelled on the virtue of being eco-friendly and thus it might first use green, renewable sources of energy and then use non-renewable sources of energy. A *deontic energy agent* might be defined as a set of rules which it must follow depending on the weather conditions, available energy and user preferences like comfort and cost. Thus, variation in the underlying models of ethics is one of the causes that lead to different agents taking the same or different actions in a specific setting.

People in different settings might have different ethical preferences. They can trust an AI-based system better, when it can act on their behalf not just functionally but also ethically [4]. AI-based systems should thus try to infer and act according to the ethical preferences of their users rather than imposing one model of ethics for everyone.

3 Representing Ethical Perspectives

To be able to reason about different ethical views and behaviour of AI, we require methods for representing these concerns mathematically and, in turn, allow for formal semantics that can be implemented in pieces of software that AI systems use or components one may deploy for (self-)governing AI behaviour. This highlights the need for ways to translate ethical concerns to mathematical/formal notions of ethics.

As discussed in the previous section, different people have different underlying paradigms of ethics based on which they make decisions. For autonomous AI-based systems to act in alignment with the ethics of the citizens they represent, it is crucial to develop computational models of different paradigms of ethics which can be incorporated into these systems. Building computational models of ethics is an active research area and it has been approached in diverse ways which will be elaborated in this section.

Machine ethics is an area which focuses on developing machines which interact with humans and other machines in an ethically acceptable way [6,7]. Top-down approaches use logical, normative or case-based reasoning; bottom-up approaches use reinforcement learning and evolutionary techniques and hybrid approaches combine both these approaches to model ethics in machines [5,52]. Artificial Moral Agents (AMAs) combine machine ethics with agency to build autonomous agents which can make ethical decisions [13,24]. Reflective equilibrium is a technique based on which agents can adjust and update their beliefs and intuitions to maximise coherence with their underlying principles [19,10]. Value Sensitive Design (VSD) is another approach which incorporates values in the system throughout the design process during conceptual, empirical and technological stages in an iterative manner [25,26].

The paradigms of ethics discussed in Section 2 have also been used in formal computational models. Utilitarianism or consequentialism has been used as the framework for building ethical robots [16,55]. Deontology has been modelled using approaches like BDI (Beliefs, Desires and Intentions) [20,37], normative models like OPF (obligated, permitted and forbidden) constraints [34], modal logic [57] and normative approaches [33]. Virtues based on the principles of virtue ethics have been modelled generally as well as in specific use cases in robots and autonomous agents [12,28,17]. Various computational approaches have also been used to compare different paradigms of ethics [14,51].

Despite the development of different ways to computationally represent ethical perspectives, there are numerous challenges which need to be tackled. In most real-world scenarios, the ethical preferences of agents are over multiple dimen-

sions which need to be properly represented. For example, in order to manage energy at home, a person might factor in their own needs, their family's needs, external climate and air quality, cost and energy preferences, based on which they regulate their home's internal control settings. A smart agent acting on behalf of a person should have data structures and algorithms to account for all these diverse parameters and then make decisions like temperature control, air purification etc. Also, ethical paradigms are relevant from different perspectives in a system– from an agent's perspective, ethical paradigms can be used to quantify if an action or choice made by the agent is ethical; on the other hand, from the systemic perspective, the choices and decisions of agents can be evaluated to check if these are aligned with the system's goals and lead the system as a whole to be in a better state.

4 Eliciting Ethical Perspectives of Individual Users

In most contexts, the ethical preferences of humans have a subjective element. Also, these are latent and sometimes it is difficult for people to convey their ethical preferences to others explicitly. For example, in a disaster response scenario, some people might want to save the most number of people, some might save the people or pets whom they know, while others might be completely indifferent. Also, the same people in a different context, say using energy in their apartment, might have completely different ethical preferences. An AI-based system which operates on behalf of users and impacts the users should not impose one ethical paradigm on all the users in the system. Rather, depending on the specific context, it should elicit, infer and then *fairly* align with the users' ethical preferences.

Also, as users interact in different settings, over time and based on their experiences, they update their ethical preferences just like other types of preferences. For example, on the road, initially, a user might be more concerned regarding efficiency and speed over caution. However, after experiencing an accident, the user might become more cautious. An agent acting on behalf of a user should be able to recognise if their ethical preferences have changed over time.

There are different approaches to eliciting the ethical preferences of users in a system. In focus groups, they can be asked to respond to questionnaires. They can be asked to act in a specific setting and their actions can be observed. They can be asked to play serious games [3] (for example the Moral Machine[3] or The Climate Game[4]). Also, LLM-based conversational interfaces combined with a knowledge base can be used to ask the users a series of questions with follow-up questions to elaborate on the underlying rationale behind their choices [61] using argumentation-based approaches [11]. Based on the users' responses, actions and choices, their ethical preferences can be inferred using these different approaches.

Some of the challenges of eliciting and inferring the ethical preferences of users are as follows. The systems are complex such that the environment and

[3] https://www.moralmachine.net/
[4] https://ig.ft.com/climate-game/

the people's preferences change over time. AI-based models should be able to dynamically adapt to these changes. People might give what they perceive to be the *right* answer instead of their *real* preferences and the system must be able to extract the real ethical preferences of users from all their responses. Also, it might be difficult to transfer these models to different settings since the users might have different ethical preferences in different contexts.

5 Aggregating Ethical Preferences

Most real-world scenarios like ridesharing, buildings, cities, country-level AI-assisted decisions, or international issues such as climate change and policy-making can be modelled as multi-agent systems with multiple agents having diverse preferences and goals. In most of these systems, the action of an AI-based application has an impact on all the stakeholders in the system. The AI-based system should be able to infer different paradigms of ethics of all the stakeholders and then aggregate their preferences to take an action which aligns with the ethical preferences of the collective. It should not act based on the ethical preferences of just one of the stakeholders but rather it should be able to correctly aggregate the ethical preferences of all the users whom it represents or are affected by its actions. Various ethical paradigms can have the following relationship– they can align, conflict or be independent with other paradigms of ethics. An ethical preference aggregation technique should be able to handle these relations among diverse models of ethics which the agents align with.

Aggregating the ethical preferences of users is relevant for autonomous agents while operating in many settings. For example, in the driving context, people have different ethical preferences regarding saving the lives of people of different age groups, professions, genders etc. [9]. An autonomous vehicle with multiple passengers having diverse preferences should be able to *fairly* infer and aggregate the preferences of all its passengers in real-time and then make decisions aligned with the ethical preferences of *all* the passengers.

Some of the challenges of aggregating multiple diverse paradigms of ethics are as follows. The ethical aggregation technique should ensure that the ethical preferences of all the involved stakeholders are accounted for. Equitability and fairness of ethical preferences of all the associated stakeholders should be ensured. In some cases, the ethical preferences of the involved users might conflict. For example, some people in an autonomous EV might prefer charging at a comparatively expensive green-energy charger while others might prefer charging at a cheaper brown-energy charger. Autonomous agents should be able to resolve such conflicts in the ethical preferences of the users whom they represent.

A lot of work has been done in the area of aggregating the choices or decisions of users. Social choice theory presents techniques to arrive at acceptable decisions representing the group specifically in social settings [46,45]. It also tries to balance the pragmatic considerations and the moral implications of the decisions. Preference aggregation involves making decisions on behalf of a group having multiple agents with diverse preferences [18]. However, preference ag-

gregation also involves challenges like handling uncertainty, incompleteness and incomparability of their preferences [56,43]. Voting-based aggregation is also a popular technique across many real-world scenarios [38]. Another similar area is judgement aggregation which uses techniques of economics, logic and computer science to aggregate individual logical judgements into a single collective judgement [22,29,23]. Argumentation-based heuristics can also be used to resolve normative conflicts among multiple agents [39].

Most of the work in the literature has been around aggregating the choices of multiple agents. As discussed above, a variety of techniques have been developed to infer and aggregate the preferences of agents. However, as discussed in Section 2, agents have diverse ethical perspectives which in turn impact their actions and choices. Aggregating ethical perspectives is characteristically different from aggregating other types of decisions or choices of agents. While preferences can be well specified, the ethical perspectives of agents are latent and abstract. Aggregating the preferences might result in one of the preferences being selected as the collective choice while aggregating the ethical perspectives of multiple agents might not be represented by one of the ethical perspectives but rather denote a new model of ethics characterising the collective. Thus, aggregating the ethical perspectives of agents is an important research direction to be explored.

6 Ensuring Alignment and Ethical Mechanism Design

AI systems which interface and impact citizen end users should align with the decisions and choices and also the ethics and values of the people they represent. Value Alignment (VA) is an active research area which focuses on ensuring that the values of AI systems are aligned with the users [47]. Value alignment involves encoding the human values and principles such that these can be incorporated in autonomous agents which can then make decisions aligned with the underlying values. It also involves exploring and evaluating different human values which will be good not just for a subset of agents but for the system as a whole [27].

For example, smart home energy management systems are used to manage non-renewable and renewable energy consumption at homes based on the preferences of the user [62,60]. So far, most of these systems have been modelled to optimise for multiple factors like cost, comfort, weather conditions and load profiles. However, there is an alternate way to model these smart home energy management systems which are aligned with the energy preferences of the users. For example, some users might prefer using green energy while others might optimise for their comfort, irrespective of the type of energy used. These systems should be able to infer the ethical preferences of users in terms of energy management and then make choices based on those preferences.

There are different ways in which value alignment can be approached and each approach has its benefits and challenges [27]. Computational models of human values have been built which can computationally model or learn different types of human values in different contexts [40]. Moreover, using preferences, values and norms have been modelled computationally [40,47]. Inverse reinforcement

learning based approaches might work in closed-world scenarios but are insufficient in realistic settings [8]. There are also hybrid approaches which combine logic and machine learning based techniques to design value-aligned systems [32].

Some of the main challenges of value alignment in AI systems are as follows. Different people have different values and these may or may not lead to the same outcomes. In this case, it is difficult to figure out which values are more suitable to be modelled in AI systems. Values are latent or abstract concepts based on which people make decisions. Thus, it is challenging to build computational models of different human values. Also, AI systems which align with some set of values should be able to demonstrate value-aligned behaviour in known and trained-for scenarios, but also in previously unseen scenarios. In case there are conflicts among different values, these systems should be able to resolve these conflicts in real-time. Also, in case no ethical option exists, the system should explain the dilemma to the user and if possible bring the system to a *safe* state.

7 Synthesis of Research Directions: A Case Discussion

We take a smart home energy management system and discuss how it can utilise the stages discussed in this paper to not just be effective, but also aligned with the ethical preferences of the people. Firstly, the system should be aware of the multiple people in the household and their possibly diverse ethical preferences *(ethics representation)*. Based on the choices and decisions they make, the system should be able to infer their individual paradigms of ethics *(ethics elicitation)*. It can also have different approaches using which it can proactively ask people for clarification or reasoning behind their actions. It should also be able to detect if the ethical preferences of the members change over time and update its behaviour accordingly. This ensures that it understands and can model the individual ethical preferences of the members. Next, if people have diverse ethical preferences regarding energy usage in the household, in should be able to aggregate them factoring everyone and ensuring that the result is fair to all *(ethics aggregation)*. In case such an action does not exist, it should notify the members with the details. Finally, the system should be such that it can be verified to ensure that its actions are aligned with the ethical preferences of the household *(ethics alignment)*. In this way, it can be ensured that an AI based system like a smart home energy management system works according to the ethical and operational preferences of the household.

8 Conclusions

Citizen-centric systems emphasise ensuring that the AI-based systems meet the needs of citizens, i.e. non-expert end users, so that citizens trust using these systems in their day-to-day lives [50]. There has been a lot of research in the area of responsible and ethical AI systems which are designed to ensure that the actions of various AI-based systems are responsible or ethical in the contexts they operate [21]. Significant work has also been done in socio-technical systems

to understand the impact of technical AI-based systems on society and vice versa [54,30,59,15]. In this paper, we focused at the intersection of these areas– responsible AI, socio-technical systems, and citizen-centric systems.

Such an ethical citizen-centric system can act as a layer on top of existing AI-based systems which operate in a socio-technical context, to ensure that the actions of these systems are ethical and aligned with the ethical preferences of the users it represents and impacts. The proposed system represents different models of ethics, elicits the ethical preferences from users, aggregates the ethical preferences of these diverse users and finally evaluates if the output of an AI-based system is aligned with the aggregated ethical output. At each step, we discuss some of the challenges which we anticipate in building such a system, as well as some possible directions and approaches which can address these challenges. Such a formulation will ensure that the actions and choices of various AI-based systems are ethically aligned with the ethical preferences of all the stakeholders in the system. This will also result in increased trust and acceptability of users towards these AI-based systems.

Acknowledgements

This work is supported by the UK Engineering and Physical Sciences Research Council (EPSRC) through a Turing AI Fellowship (EP/V022067/1) on Citizen-Centric AI Systems (https://ccais.ac.uk/) and by Responsible Ai UK (EP/Y009800/1) (https://rai.ac.uk/). We also thank the PRICAI reviewers for their constructive feedback. For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) licence to any author-accepted manuscript version arising.

References

1. Round table discussion during RAI UK Partner Network Town Hall (2024)
2. Round table discussion during Responsible AI Community Building Event (2024)
3. Abt, C.C.: Serious games. University Press of America (1987)
4. Alaieri, F., Vellino, A.: Ethical decision making in robots: Autonomy, trust and responsibility. In: Social Robotics: ICSR. pp. 159–168. Springer (2016)
5. Allen, C., Smit, I., Wallach, W.: Artificial morality: Top-down, bottom-up, and hybrid approaches. Ethics and information technology **7**, 149–155 (2005)
6. Anderson, M., Anderson, S.L.: Machine ethics: Creating an ethical intelligent agent. AI magazine **28**(4), 15–15 (2007)
7. Anderson, M., Anderson, S.L.: Machine ethics. Cambridge University Press (2011)
8. Arnold, T., Kasenberg, D., Scheutz, M.: Value alignment or misalignment–what will keep systems accountable? In: Workshops at AAAI-2017 (2017)
9. Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.F., Rahwan, I.: The moral machine experiment. Nature **563**(7729), 59–64 (2018)
10. Beisbart, C., Betz, G., Brun, G.: Making reflective equilibrium precise. a formal model. Ergo: an open access journal of philosophy **8**(15), 441–472 (2021)
11. Bench-Capon, T.J.: Persuasion in practical argument using value-based argumentation frameworks. Journal of Logic and Computation **13**(3), 429–448 (2003)

12. Berberich, N., Diepold, K.: The virtuous machine-old ethics for new technology? arXiv preprint arXiv:1806.10322 (2018)
13. Cervantes, J.A., López, S., Rodríguez, L.F., Cervantes, S., Cervantes, F., Ramos, F.: Artificial moral agents: A survey of the current status. Science and engineering ethics **26**(2), 501–532 (2020)
14. Chhabra, J., Sama, K., Deshmukh, J., Srinivasa, S.: Evaluating computational models of ethics for autonomous decision making. AI and Ethics pp. 1–14 (2024)
15. Chopra, A.K., Singh, M.P.: Sociotechnical systems and ethics in the large. In: Proceedings of the 2018 AAAI/ACM Conference on AIES. pp. 48–53 (2018)
16. Cloos, C.: The utilibot project: An autonomous mobile robot based on utilitarianism. In: 2005 AAAI Fall Symposium on Machine Ethics. pp. 38–45 (2005)
17. Coleman, K.G.: Android arete: Toward a virtue ethic for computational agents. Ethics and Information Technology **3**(4), 247–265 (2001)
18. Conitzer, V.: Computational aspects of preference aggregation. Ph.D. thesis, Carnegie Mellon University (2006)
19. Daniels, N.: Justice and justification: Reflective equilibrium in theory and practice. Cambridge University Press (1996)
20. Dennis, L., Fisher, M., Slavkovik, M., Webster, M.: Formal verification of ethical choices in autonomous systems. Robotics and Autonomous Systems (2016)
21. Dignum, V.: Responsible artificial intelligence: how to develop and use AI in a responsible way, vol. 1. Springer (2019)
22. Endriss, U.: Judgment aggregation (2016)
23. Fioravanti, F., Rahwan, I., Tohmé, F.A.: Classes of aggregation rules for ethical decision making in automated systems. arXiv preprint arXiv:2206.05160 (2022)
24. Formosa, P., Ryan, M.: Making moral machines: why we need artificial moral agents. AI & society **36**(3), 839–851 (2021)
25. Friedman, B.: Value-sensitive design. interactions (1996)
26. Friedman, B., Hendry, D.G., Borning, A., et al.: A survey of value sensitive design methods. Foundations and Trends in Human–Computer Interaction (2017)
27. Gabriel, I.: Artificial intelligence, values, and alignment. Minds and machines **30**(3), 411–437 (2020)
28. Govindarajulu, N.S., Bringsjord, S., Ghosh, R., Sarathy, V.: Toward the engineering of virtuous machines. In: Proceedings of AI, Ethics, and Society (2019)
29. Grossi, D., Pigozzi, G.: Judgment aggregation: a primer. Springer Nature (2022)
30. Jones, A.J., Artikis, A., Pitt, J.: The design of intelligent socio-technical systems. Artificial Intelligence Review **39**, 5–20 (2013)
31. Kant, I.: Moral law: Groundwork of the metaphysics of morals. Routledge (2013)
32. Kim, T.W., Hooker, J., Donaldson, T.: Taking principles seriously: A hybrid approach to value alignment in artificial intelligence. JAIR **70**, 871–890 (2021)
33. Luck, M., Mahmoud, S., Meneguzzi, F., Kollingbaum, M., Norman, T.J., Criado, N., Fagundes, M.S.: Normative agents. Agreement technologies pp. 209–220 (2013)
34. Malle, B.F., Scheutz, M., Austerweil, J.L.: Networks of social and moral norms in human and robot agents. In: A world with robots, pp. 3–17. Springer (2017)
35. Mill, J.S.: Utilitarianism. In: Seven masterpieces of philosophy, pp. 329–375. Routledge (2016)
36. Mill, J.S., Bentham, J.: Utilitarianism and other essays. Penguin UK (1987)
37. Neto, B.d.S., da Silva, V.T., de Lucena, C.J.: Nbdi: An architecture for goal oriented normative agents. In: ICAART 2011 (2011)
38. Noothigattu, R., Gaikwad, S., Awad, E., Dsouza, S., Rahwan, I., Ravikumar, P., Procaccia, A.: A voting-based system for ethical decision making. In: AAAI (2018)

39. Oren, N., Luck, M., Norman, T.J.: Argumentation for normative reasoning. In: Proc. Symp. Behaviour Regulation in Multi-Agent Systems. pp. 55–60 (2008)
40. Osman, N., d'Inverno, M.: A computational framework of human values. In: AAMAS-24 (2024)
41. Pěchouček, M., Mařík, V.: Industrial deployment of multi-agent technologies: review and selected case studies. AAMAS **17**(3), 397–431 (2008)
42. Pechoucek, M., Thompson, S.G., Voos, H.: Defence Industry Applications of Autonomous Agents and Multi-Agent Systems. Springer (2008)
43. Pini, M.S., Rossi, F., Venable, K.B., Walsh, T.: Incompleteness and incomparability in preference aggregation: Complexity results. Artificial Intelligence **175**(7-8), 1272–1289 (2011)
44. Russell, S.: Human compatible: AI and the problem of control. Penguin Uk (2019)
45. Sen, A.: Social choice theory: A re-examination. Econometrica pp. 53–89 (1977)
46. Sen, A.: Social choice theory. Handbook of mathematical economics **3**, 1073–1181 (1986)
47. Sierra, C., Osman, N., Noriega, P., Sabater-Mir, J., Perelló, A.: Value alignment: a formal approach. arXiv preprint arXiv:2110.09240 (2021)
48. Solomon, W.D.: Normative ethical theories. Ch. K. Wilber, Economics, ethics and public policy, Boston, Rowman & Littlefield Publishers pp. 119–138 (1998)
49. Steen, M.: Ethics as a participatory and iterative process. Communications of the ACM **66**(5), 27–29 (2023)
50. Stein, S., Yazdanpanah, V.: Citizen-centric multiagent systems. In: AAMAS 2023. pp. 1802–1807 (2023)
51. Tennant, E., Hailes, S., Musolesi, M.: Modeling moral choices in social dilemmas with multi-agent reinforcement learning. In: IJCAI-2023. pp. 317–325 (2023)
52. Tolmeijer, S., Kneer, M., Sarasua, C., Christen, M., Bernstein, A.: Implementations in machine ethics: A survey. ACM Computing Surveys (CSUR) **53**(6), 1–38 (2020)
53. Trianosky, G.: What is virtue ethics all about? American Philosophical Quarterly **27**(4), 335–344 (1990)
54. Van Dam, K.H., Nikolic, I., Lukszo, Z.: Agent-based modelling of socio-technical systems, vol. 9. Springer Science & Business Media (2012)
55. Van Dang, C., Tran, T.T., Gil, K.J., Shin, Y.B., Choi, J.W., Park, G.S., Kim, J.W.: Application of soar cognitive agent based on utilitarian ethics theory for home service robots. In: URAI. pp. 155–158. IEEE (2017)
56. Walsh, T.: Uncertainty in preference elicitation and aggregation. In: AAAI. vol. 7, pp. 3–8 (2007)
57. Wiegel, V., van den Berg, J.: Combining moral theory, modal logic and mas to create well-behaving artificial agents. International Journal of Social Robotics **1**(3), 233–242 (2009)
58. Winfield, A.F., Jirotka, M.: Ethical governance is essential to building trust in robotics and artificial intelligence systems. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences (2018)
59. Woodgate, J.M., Ajmeri, N.: Macro ethics for governing equitable sociotechnical systems. In: AAMAS'22. pp. 1824–1828 (2022)
60. Zafar, U., Bayhan, S., Sanfilippo, A.: Home energy management system concepts, configurations, and technologies for the smart grid. IEEE access (2020)
61. Zheng, L., Chiang, W.L., Sheng, Y., et al.: Judging llm-as-a-judge with mt-bench and chatbot arena. Advances in Neural Information Processing Systems **36** (2024)
62. Zhou, B., Li, W., Chan, K.W., Cao, Y., Kuang, Y., Liu, X., Wang, X.: Smart home energy management systems: Concept, configurations, and scheduling strategies. Renewable and Sustainable Energy Reviews **61**, 30–40 (2016)