Bickley, Steve J.; Torgler, Benno

Working Paper

Cognitive architectures for artificial intelligence ethics

CREMA Working Paper, No. 2021-27

Provided in Cooperation with:
CREMA - Center for Research in Economics, Management and the Arts, Zürich

Suggested Citation: Bickley, Steve J.; Torgler, Benno (2021) : Cognitive architectures for artificial intelligence ethics, CREMA Working Paper, No. 2021-27, Center for Research in Economics, Management and the Arts (CREMA), Zürich

This Version is available at:
https://hdl.handle.net/10419/246010

CREMA

Center for Research in Economics, Management and the Arts

Cognitive Architectures for Artificial Intelligence Ethics

Working Paper No. 2021-27

CREMA Südstrasse 11 CH - 8008 Zürich www.crema-research.ch

Cognitive Architectures for Artificial Intelligence Ethics

Steve J. Bickley[1,2*] and Benno Torgler[1,2,3]

[1]School of Economics and Finance, Queensland University of Technology

[2]Centre for Behavioural Economics, Society and Technology, Australia

[3]Centre for Research in Economics, Management, and the Arts (CREMA), Switzerland

Abstract: *As artificial intelligence (AI) thrives and propagates through modern life, a key question to ask is how to include humans in future AI? Despite human-involvement at every stage of the production process from conception and design through to implementation, modern AI is still often criticized for its "black box" characteristics. Sometimes, we do not know what really goes on inside or how and why certain conclusions are met. Future AI will face many dilemmas and ethical issues unforeseen by their creators beyond those commonly discussed (e.g., trolley problems and variants of it) and to which solutions cannot be hard-coded and are often still up for debate. Given the sensitivity of such social and ethical dilemmas and the implications of these for human society at large, when and if our AI make the "wrong" choice we need to understand how they got there in order to make corrections and prevent recurrences. This is particularly true in situations where human livelihoods are at stake (e.g., health, well-being, finance, law) or when major individual or household decisions are taken. Doing so requires opening up the "black box" of AI; especially as they act, interact, and adapt in a human world and how they interact with other AI in this world. In this article, we argue for the application of cognitive architectures for ethical AI. In particular, for their potential contributions to AI transparency, explainability, and accountability. We need to understand how our AI get to the solutions they do, and we should seek to do this on a deeper level in terms of the machine-equivalents of motivations, attitudes, values, and so on. The path to future AI is long and winding but it could arrive faster than we think. In order to harness the positive potential outcomes of AI for humans and society (and avoid the negatives), we need to understand AI more fully in the first place and we expect this will simultaneously contribute towards greater understanding of their human counterparts also.*

Keywords Artificial Intelligence, Ethics, Cognitive Architectures, Intelligent Systems, Ethical AI, Society

* Corresponding author (steven.bickley@hdr.qut.edu.au).

> Over time, as we recognize more consequences of our actions, our societies tend to give us both responsibility and accountability for these consequences—credit and blame depending on whether the consequences are positive or negative. Artificial intelligence only changes our responsibility as a special case of changing every other part of our social behaviour… [However], [r]esponsibility is not a fact of nature. Rather, the problem of governance is as always to design our artifacts—including the law itself—in a way that helps us maintain enough social order so that we can sustain human dignity and flourishing.
> (Bryson, 2020, p. 5)

> To my mind, progress on giving computers moral intelligence cannot be separated from progress on other kinds of intelligence; the true challenge is to create machines that can actually *understand* the situation that they confront. As Isaac Asimov's stories demonstrate, a robot can't reliably follow an order to avoid harming a human unless it can understand the concept of *harm* in different situations. Reasoning about morality requires one to recognize cause-and-effect relationships, to imagine different possible futures, to have a sense of the beliefs and goals of others, and to predict the likely outcomes of one's actions in whatever situation one finds oneself. In other words, a prerequisite to trustworthy moral reasoning is general common sense, which, as we've seen, is missing in even the best of today's AI systems.
> (Mitchell, 2019, p. 156-157).

1 Introduction

As artificial intelligence (AI) thrives and propagates through modern life changing our society as electricity changed our society a century ago[1], a key question to ask is *how to include humans in future AI?* In the age of (vividly) personalised web search results and social media feeds, AI clearly informs to a non-insignificant degree what information we receive about the world around us in a sort of 'filter bubble' (Pariser, 2011), automatically choosing what is most relevant and interesting to us on our behalf. While useful in some contexts (e.g., automated movie and song suggestions), this gives AI the power to choose what we see and hear about the world or more appropriately – *our world* – the one that has been created for us and communicated to us by AI, often largely without any input on our own part. Further, as "we are exposed to certain kinds of stimuli" AI can "learn how we respond to them and how these stimuli can be used to trigger certain behavioural responses" (Helbing, 2019, p. 28), tiptoeing eerily close to an infringement of free will and self-determination as if we would each be part of a gigantic Skinner box. While AI already brings and will bring a lot more social and societal benefits, AI systems can be designed to manipulate our decisions and behaviour towards a specific agenda, using methods linked to messages and values we are most susceptible and sensitive to. This of course can be used for good, but in the wrong hands it obviously causes concerns. In general, technological developments often start with good intentions behind them.

[1] See Andrew Ng: Why AI Is the New Electricity | Stanford Graduate School of Business

However, the line between 'good' and 'bad' are something early adopters define for themselves and can lead them to walk the grey areas in between. As do the notions behind 'big nudging' and 'citizen scores' developed and advanced by data-driven cybernetic societies such as Singapore and China (Helbing, 2019). As opposed to the idea of libertarian paternalism that emphasizes the importance of freedom of choice and the goal to influence people's behaviour in order to make their lives better, healthier, and therefore longer (Thaler and Sunstein, 2009). It is still early days for AI and so, we need to be prepared for the directions which future AI may take, and the implications of these in a social, health and well-being, and moral context.

Despite increasing regulations, many critics cite the 'black box' characteristics of AI systems, models, and behaviours as problematic for existing and future AI (Carabantes, 2020). This is despite human-involvement at every stage of the production process, from conception and design through to implementation. Sometimes, we do not know what really goes on inside or how and why certain conclusions are met. If we do not understand them*, should we let them reign so free and pervasive through human life and society?* Human life is complex, messy, and unpredictable. Something humans do have is intentionality; reasons for what they do and why they do it (Searle & Willis, 1983). We share this expectation of intentionality with others (Malle & Knobe, 1997) and we will come to expect this of future AI also. Explainable AI (X-AI) and transparent reasoning aims to aid in our ability to understand and communicate how an AI system or model makes it decisions in clear and coherent ways (Gunning et al., 2019). For example, X-AI requires that "[w]hen asked a question about its activities, the agent must be able to retrieve the ways in which its choices relate to norms and then communicate them in accessible terms" (Langley, 2019, p. 9778). X-AI also prompts the human designer and user to reflect on their own knowledge, biases, and possible (mis)conceptions as they make sense of the AI's reasoning and naturally, compare it to their own via introspection (Richards, 2019) and counterfactuals (Costello & McCarthy, 1999) (i.e., imagined *what-if* scenarios, situations, and non-experiences). This allows also in some sense a forecast for (un)intended consequences, particularly useful in restoration/conservation settings or where payoffs are realised in the distant future (Mozelewski & Scheller, 2021). Further, this allows an opportunity to implement the necessary controls *before* systems become live. Reflection is a powerful tool which gives us opportunity to adapt to changing situations by thinking about outcomes achieved and how we got there. In the process, we change the way we think (i.e., goals, intentions, motivations) and act and build on our tacit, learned, and experience-based knowledge base. It also allows us to explain and make sense of our reasoning and further, to be transparent when communicating

to others the rules and procedures we apply to get there. This is an important factor in trust development (Chen & Barnes, 2014), and an indicator of human (mis)use of automation at the human-machine interface (Chen & Barnes, 2014; Visser et al., 2018), i.e., when being operated directly by the end-user. The lack of transparency makes AI systems more vulnerable and potentially subject to sabotage and misuse. Mitchell (2019), for example, stresses:

> Machine learning is being deployed to make decisions affecting the lives of humans in many domains. What assurances do you have that the machine creating your news feed, diagnosing your diseases, evaluating your loan applications, or – God forbid – recommending your prison sentence have learned enough to be trustworthy decision makers? (p. 142).

A focus on cognitive architectures can help increase procedural transparency, reducing sabotage and misuse (e.g., via introducing common sense) and help to move towards a better understanding of how to model aspects such as emotions, well-being, or empathy. Cognitive architectures allow to answer *why* questions and put weight on the ability to envision alternative options and realities (counterfactuals) and compare them. To interpret data also means to formulate a model of the data generating process and reflect on actions taken or not taken. Cognitive architectures can help navigate in a world rich in causal and unpredictable forces which is a challenge when applying purely a machine learning approach:

> Like the prisoners in Plato's famous cave, deep-learning systems explore the shadows on the cave wall and learn to accurately predict their movements. They lack the understanding that the observed shadows are mere projections of three-dimensional objects moving in a three-dimensional space. Strong AI requires this understanding" (Pearl and Mackenzie, 2018, p. 362).

Pearl and Mackenzie (2018) argue that an AI system would require a causal model of the world and a causal model of its own software (reflect on its own actions). In addition, a memory that records how intents in its mind are connected to events in the outside world (p. 367). This echoes calls made more recently by brain scientist, Jeff Hawkins, in his book *A Thousand Brains: A New Theory of Intelligence*. Hawkins (2021) argues the mind is constantly creating and revising its models of the world and the objects in it based on how we interact with them. As new problems call for new solutions, additional cognitive functions emerge to offer a path forward due their flexibility. Necessarily, this requires the ability to think about our thinking (i.e., meta-cognition), revise what we currently know to be true based on new evidence (i.e., learning and extension (not replacement) of existing knowledge), and to imagine and choose between the potential consequences of various actions which offer solutions to the problem and situation at hand.

A focus on cognitive architectures is therefore useful also from a societal perspective as a better procedural understanding can help increase trust in AI systems via a better understanding and interpretation and improved assignment of accountability. Understanding cognitive architectures are also important as AI systems are still quite limited relative to what human intelligence can achieve and certainly fall short of the expectations of artificial *general* intelligence (AGI). We argue that to find answers to the core AI questions that Mitchell (2019) classifies the "Great AI Trade-Off" requires a better understanding of cognitive architectures:

> Should we embrace the abilities of AI systems, which can improve our lives, and allow these systems to be employed ever more extensively? Or should we be more cautious, given current AI's unpredictable errors, susceptibility to bias, vulnerability to hacking and lack of transparency in decision-making? To what extent should humans be required to remain in the loop in different AI applications? What should we require of an AI system in order to trust it enough to let it work autonomously?

Cognitive architectures are conceptual models of intelligent minds – be it human, animal, or artificial – as they learn, process, store, and reuse knowledge and information, and make and carry out decisions to problems they face. Cognitive architectures can help implement transparency and explainability in AI with different levels or subsystems each performing distinct yet interrelated cognitive functions including value and goal setting, planning, deliberation, and action to name a few – particularly in development of artificial moral agents (Cervantes et al., 2020). Cognitive architectures also improve the ability to communicate to a wider audience beyond experts and specialists in academia and government and interact with them socially (Samsonovich, 2020). It provides clear frameworks that are flexible to the needs of the user, architect, and environment of application. Further, it helps clarify the assignment of responsibility across different levels of a multi-level cognitive framework, lending itself to division of labour and division of credit and blame (i.e., accountability). However, as Griffiths and Lucas (2016) contend, everyday "[e]conomic transactions, like legal transactions, do not take place in a vacuum, but in a social and moral context" (p. 30). This speaks to the importance of social and moral factors in everyday transactions that AI, as humans have, will likely face on a regular day-to-day basis. Further, these transactions are by "human beings with a 'mindset' of motivations and aspirations which determine how they react to the particularity of the time and circumstances in which they find themselves" (Griffiths and Lucas, 2016, p. 30). This shows the (p)relevance of goals, aspirations, and motivations in human life and society and also requires AI reasoning about others' mental models and states – another support for

cognitive architectures in ethical AI. Doing so requires opening up the "black box" of AI; especially as they act, interact, and adapt in a human world and how they interact with other AI in this world.

In this contribution, we argue for the application of cognitive architectures for AI ethics. In particular, for their potential contributions to AI transparency, explainability, and accountability. Future AI will face many dilemmas and ethical issues unforeseen by their creators beyond those commonly discussed (e.g., trolley problems and variants of it) and to which solutions cannot be hard-coded and are often still up for debate. Given the sensitivity of such social and ethical dilemmas and the implications of these for human society at large, when and if our AI make the "wrong" choice we need to understand how they got there in order to make corrections and prevent recurrences, and if necessary, punish or reward. This is particularly true in situations where human livelihoods are at stake (e.g., health, well-being, finance, law) or when major individual or household decisions are taken. In the next section, we introduce AI ethics as they apply to society, namely: ethics in design, ethics by design, and ethics for design (Dignum, 2019). Following this, we discuss cognitive architectures as they apply to intelligent minds in individuals and collectives. We then discuss their relevance to AI ethics and their potential contributions to transparency, explainability, and accountability. Finally, we summarise with implications, shortcomings and challenges, and future perspectives for cognitive architectures in AI ethics and future AI more generally.

2 AI Ethics

Human life is full of moral and ethical dilemmas and as such, plenty of opportunity to practice our moral reasoning and ethical decision-making. These transactions "do not take place in a vacuum, but in a social and moral context" (Griffiths and Lucas, 2016, p. 30) and hence, an abstraction or idealisation that is void of historical time, and space makes little sense. Should Odysseus risk encountering Charybdis and lose his entire ship and companions, or should he sail closer to the evil monster Scylla which would lead in losing six of his crew members:

> We then sailed on up the narrow strait with wailing. For on one side lay Scylla and on the other divine Charybdis terribly. We then sailed on up the narrow strait with wailing. For on one side lay Scylla and on the other divine Charybdis terribly sucked down the salt water of the sea. Verily whenever she belched it forth, like a cauldron on a great fire she would seethe and bubble in utter turmoil, and high over head the spray would fall on the tops of both the cliffs. But as often as she sucked down the salt water of the sea, within she could all be seen in utter turmoil, and round about the rock roared terribly, while beneath the earth appeared black with sand; and

> pale fear seized my men. So we looked toward her and feared destruction; but meanwhile Scylla seized from out the hollow ship six of my comrades who were the best in strength and in might (Homer 1945, The Odyssey, p. 449).

When choosing what to do we rely on our morals. We also subscribe to shared ethical principles with those we interact with frequently and have relationships with. Morals (a.k.a. moral values) are values (i.e., codes, standards of practice, guiding principles) that protect and enhance life; for all, self *and* others. Being moral means knowing the difference between right and wrong and wanting/choosing to do what is right. Of course, what is right and wrong is subjective; a matter of opinion and preference and is often driven by culture, religion, and environment (Awad et al., 2018) among other contextual factors. Morality also develops and evolves over a lifetime (Brady and Hart, 2007); from focusing on personal interests through to maintaining social norms and then more independent critical thinking about morals and ethics. Being ethical means carrying out morals, typically those shared with other individuals, groups, and institutions. In other words, ethics are those morals which emerge from shared understanding and agreement and are often tied to a specific socio-political environment, time and place. They are often used to foster safety, security, and cooperation in communities of individuals who interact frequently. However, being ethical does not always equate to being moral. For example, adherence to the criminal's code of silence – designed to protect criminals from police conviction – is ethical behaviour from the standpoint of other criminals. It is viewed favourably upon by other criminals and rejection of this code often leads to serious consequences ("snitches get stitches"). Morally speaking, lying and misleading are morally inadmissible acts as is criminal activity in the first place. Hence, sometimes (often) ethics and morals collide.

As AI proliferates further through human life and takes on increasingly difficult tasks in more complex environments, the likelihood that AI will face problems and events with moral and ethical implications will increase dramatically. This is particularly true where human life and/or livelihoods are at stake and when big government, organisational, household, or individual decisions are made. This is where AI ethics has a solid role to play. Broadly speaking, the AI ethics literature can be broadly clustered by the ethics of AI (i.e., ethical issues related to or caused by AI) and ethical AI (i.e., machine ethics / the ethical and moral behaviour of AI) (Siau and Wang, 2020). Essentially, this boils down to the *What* and *How* of AI. In other words, what are the effects of AI on society (and vice-versa), and how can we ensure AI act ethically by design, monitoring, and regulation. Whilst this is a fairly intuitive (dualistic) distinction, we

prefer the *responsibility* approach (i.e., implementation-focus) of Dignum (2019) who clusters AI ethics into three areas of focus:

1 *Ethics in Design* (i.e., the regulatory/engineering processes that support design and evaluation of AI as it applies to societal interests).
2 *Ethics by Design* (i.e., the ethical behaviour of AI, a.k.a. ethical AI).
3 *Ethics for Design* (i.e., the codes of conducts, standards, and regulations and certification processes for AI research, design, construction, use, operation and maintenance, and decommissioning).

This provides a mean to discuss what ethical AI should or ought to look like, the roles and functions they should perform in society, and how this can be practically realised. Further, what the role of regulators should be and how should we balance ethical, legal, economic, and social considerations to achieve something which benefits society as a whole. Clearly, the AI ethics literature is broad and varied. For example, Jobin et al. (2019) in their review of 84 AI ethics soft-law and grey literature find eleven overarching ethical values and principles (by frequency of appearance): transparency, justice and fairness, non-maleficence, responsibility, privacy, beneficence, freedom and autonomy, trust, dignity, sustainability, and solidarity. The core ethical principles which featured in more than half of the sources include transparency, justice and fairness, non-maleficence, responsibility, and privacy. AI4People[2] also include beneficence and autonomy and highlight the salience of explicability in enabling the application of ethics principles to AI (Floridi et al., 2018). For example, they highlight "… for AI to be beneficent and non-maleficent, we must be able to understand the good or harm it is actually doing to society" (p. 700). In this contribution, we focus on explicability, which is itself backed by transparency and explainability and itself lends support to accountability. This stems from "… the need to understand and hold to account the decision-making processes of AI" (Floridi et al., 2018, p. 700) and to ensure the people and companies responsible for developing and deploying the AI are also held accountable in the case of negative outcomes for individuals and society. We discuss the relevance of cognitive architectures to this cause later in Section 4.

In the following subsections, we introduce and unpack each of Dignum's (2019) AI Ethics clusters in turn. In particular, we focus on the AI ethics principles of transparency, explainability, and accountability whose presence is felt right across these three domains of AI

[2] An international scientific committee focused on developing *Good AI Society,* https://www.eismd.eu/ai4people/

ethics and for the reasons laid out above. In other words, we see accountability to be critical to the sustainability of AI (longer term implications in fostering safety, security, and accountability across every stage of the AI production process). To this end, transparency and explainability are required for policymaking, community consultation, layman communication and understanding, interdisciplinary knowledge sharing, and monitoring adherence to AI rules and regulations.

2.1 Ethics in Design

Ethics in Design refers to the regulatory and engineering processes which support the design and evaluation of AI systems as they integrate into modern society (Dignum, 2019). It is about ensuring the critical evaluation of social, legal, and ethical implications of AI as they transform more traditional (socio)economic systems and structures. Essentially, it is about leveraging the beneficial outcomes of AI and avoiding the negative by way of due diligence and critical thinking. As stressed, this relies heavily on the principles of *explainability* (i.e., ability of AI to explain rationale behind its decisions and behaviours and to explain its reasoning and assumptions), *accountability* (i.e., the role of people as they develop, manufacture, sell, and use AI systems including considerations of liability, autonomy/oversight, and legal/regulatory requirements), and *transparency* (i.e., openness about data, design processes, algorithms, and with societal actors and stakeholders). These principles allow and encourage informed participation by a diverse group of stakeholders (e.g., researchers, citizens, policymakers) and are necessary for relevant discussions and debate to take place. They also facilitate law and governance throughout the AI production process by assigning *liability across the full lifecycle*, *making explicit the processes and assumptions* used to make decisions and form conclusions, and give *documented reporting of the AI development processes* leveraging existing best practices in software engineering for stakeholder engagement, version control, verification and validation testing, among others. In taking these discussions to a wider audience, we can decide together what are the appropriate forms of conduct for AI and their makers, what the extent of AI autonomy and its pervasiveness in modern society should be, and how we should look to ensure *AI for All* and *AI for Good*[3].

van den Hoven et al. (2015) provide historical account of including social and moral values in the technology design/development process; often referred to as Value Sensitive Design (VSD) or *Design for Values*. By focusing on the design process, moral and societal considerations can

[3] See, for example, Berendt's (2019) discussion of AI for Common Good.

be incorporated from the ground up; informed by the context of application and guiding the design towards more sustainable outcomes. However, this requires meeting moral expectations on top of functional design requirements, a more difficult task. In addition, a *Responsible Research and Innovation (RRI)* approach (Owen et al., 2012) can help foster engagement with all societal stakeholders during the research and innovation process, right up until the introduction of resulting AI products and services into the market. RRI rests on addressing four key AI issues (Dignum, 2019): *openness and transparency* (e.g., of funding, decision-making processes), *anticipation and reflection* (e.g., preventive risk management of social, environmental, and economic long- and short-term impacts), *responsiveness and adaptability* (e.g., ability to respond to changing circumstances, norms, and societal expectations), and *diversity and inclusion* (e.g., engagement of diverse stakeholder groups in design and development process). Together, VSD and RRI (and their variants) provide a good starting place for AI *Ethics in Design*. The key takeaway being an open and documented design process which includes deliberation[4] and active engagement of diverse stakeholder groups with varying expertise, background/biases, and perspectives to better align AI systems with the needs, values, and expectations of society as a whole.

2.2 Ethics by Design

Ethics by Design refers to the ethical behaviour of AI and the technical means for achieving this (Dignum, 2019). In other words, the integration of ethical reasoning abilities and building of safeguards in the design and development of AI systems. Essentially, this requires aligning human and AI values and giving AI the means to behave and reason ethically. The increasing (p)relevance of AI in modern life begs the question as to whether we should embed AI with ethical values and moral guides (van de Poel, 2020). If so, which ones should we embed and how?[5] Machine learning approaches to moral and ethical decision-making based on human-labelled data may at best mimic the morality of humans. Humans are themselves susceptible to bias, mis-guided thinking and decision making, and making errors so maybe we should aim to do better than humans. This then begs to ask whether AI should have certain rights and moral status itself (Müller, 2021). If so, which AI, under what circumstances, and why? This will become increasingly relevant as people tend to anthropomorphise AI systems and in turn,

[4] Legitimate deliberation being underpinned by five essential characteristics (Fishkin, 2011): information (i.e., accurate and relevant data), substantive balance (i.e., balanced evidence base), diversity (i.e., participation by all relevant stakeholders), conscientiousness (i.e., integrity in evaluation process), and equal consideration (i.e., evidence-based decision making).

[5] For example, the top-down (logic and theory-based), bottom-up (adaptive, learned, and data-driven), and hybrid approaches to AI ethical and moral reasoning (Wallach & Allen, 2008; Dignum, 2019).

assign ‘human’ traits and characteristics to AI. This conceptualisation also permeates the research community, but it can be misleading (Salles et al., 2020). Further, it is likely we expect more of those rights (ethically speaking) as AI become more autonomous and socially aware (and involved) (Wallach & Allen, 2008). As AI becomes increasing human-like we as a society may increasingly come to view them as such and in turn, may ask e.g., whether it is ethically admissible to have them performing boring or dangerous work on behalf of humans in the first place. Even under a human-centred design approach, we can maintain that it is unethical to intentionally cause harm to an AI “because even if mutilating a robot does not harm the robot (because the robot is not the kind of thing that can be harmed), such mutilation may in fact do harm to the humans involved… [t]he idea is that even if robots cannot be harmed, they are, at least sometimes, ‘made in our image’ to such an extent that wilfully abusing them is at best grotesque, at worst unethical” (Chrisley, 2020, p. 468). At times such recursive thinking can become overwhelming and seems more relevant only very far in the future (i.e., future AI). However, a proactive approach to potential moral and ethical issues of future AI allow us to address them early on before they become too entrenched and cumbersome to address and unpack later on (Collingridge, 1980).

Moor (2006) makes the distinction between implicit and explicit ethical AI. Implicit ethical AI are those built-in ethical features that promote (avoid) ethical (unethical) behaviour from occurring in the first place. For example, collision-avoidance in self-driving cars serves to safely deliver its passengers to their destination – an implicit promise made between passenger and driver. In contrast, explicit ethical AI provides explicitly the tools and ability to reason about ethical information and decide what course(s) of action (or inaction) may be most appropriate (ethically speaking) in any given situation. In addition, explicit AI could sometimes violate certain rules to better meet overarching ethical obligations (Bench-Capon and Modgil, 2017). For example, crossing-over to drive on the wrong side of the road (breaking road traffic rules) to avoid a collision with a pedestrian (to protect human life) would be acceptable where it causes no additional harm to others (e.g., other road users). Whilst it may at first seem chaotic that AI are free to choose which rules to follow and when, it allows for much more complex and adaptive behaviour that are responsive to the current situation and context. This adaptivity is important for a constantly changing world where new and novel problems, challenges, and opportunities continually arise leading to many situations unforeseen by AI designers. Of course, explicit ethical AI could present challenges for existing legal frameworks which appear more applicable to cases involving implicit ethical AI (Dyrkolbotn et al., 2018). Here, value

alignment is important and needs to allow for a broad class of different users, problems, and contexts. The VSD and RRI approaches (described Section 2.1) are again useful here in identifying and engaging with the most relevant societal stakeholders and purposefully integrating their aggregated views in meaningful and inclusive ways.

2.3 Ethics for Design

Ethics for Design focuses on the practical requirements to ensure the integrity of those who research, design, develop, deploy, and manage/maintain AI systems. This includes codes of conduct, regulatory requirements, industrial standards, and certification processes (Dignum, 2019). In other words, the documented processes and requirements which provide specific advice and guidance for achieving ethical AI, that allow traceability through AI development, and that demonstrate that risks have been systematically identified and controls for these introduced to reduce risk likelihoods to as low as reasonably practicable. Essentially, this is about ensuring ethical AI in practice and provides AI designers, developers, and organisations with the actionable tools and information they need to achieve ethically sound AI and outcomes for society. Further, "deciding on ethical guidelines, governance policies, incentives and regulations" (Dignum, 2019, p. 94) and certification and monitoring of these. Work on such standardisation has already begun with IEEE's Ethically Aligned Design series[6] and deliberation at international platforms such as the AI4People Summit[7] and the Asilomar Conference on Beneficial AI[8] in 2017 and AI Safety[9] in 2015. Research institutes such as the Future of Humanity Institute[10], Future of Life Institute[11], Leverhulme Centre for the Future of Intelligence (CFI)[12], and Centre for the Study of Existential Risk[13] tackle many of the big and longer-term AI problems from future of work to individual privacy and autonomy. The Partnership on AI[14] provides another such cooperative extending across academia and industry, providing an open platform for discussion and engagement about AI and its influences on people and society. Essentially, this area of AI ethics seeks to define the rulebook for those who develop, manufacture, implement, and maintain AI systems and define also how this will be implemented and overseen (e.g., monitoring, regulation, legislation) and by whom (e.g., AI

[6] IEEE Ethically Aligned Design series: https://ethicsinaction.ieee.org/
[7] AI4 People Summit: https://www.eismd.eu/ai4people/
[8] Asilomar Conference on Beneficial AI: https://futureoflife.org/bai-2017/
[9] Asilomar Conference on AI Safety: https://futureoflife.org/2015/10/12/ai-safety-conference-in-puerto-rico/
[10] Future of Humanity Institute: https://www.fhi.ox.ac.uk/
[11] Future of Life Institute: https://futureoflife.org/
[12] Leverhulme Centre for the Future of Intelligence (CFI): http://lcfi.ac.uk/
[13] Centre for the Study of Existential Risk: https://www.cser.ac.uk/
[14] Partnership on AI: https://www.partnershiponai.org/

ethical watchdog). The rulebook will also vary for different industries rather than some more generic regulatory approach. For a human-centred AI approach, Shneiderman (2020) provides a three-level governance structure incorporating *team* technical practices and software engineering, *organisation* management strategies and standards, and *industry* oversight, regulation, and policymaking. A contextual morality framework shows promising results for representing the diversity of viewpoints, backgrounds, and environmental constraints in AI systems (van Berkel et al., 2020).

There is an increasing need to understand how decisions are made by AI methods, particularly when these decisions affect humans' lives in non-insignificant ways (Goodman & Flaxman, 2017). For example, in the areas of health, well-being, finance, and law. This concern is made explicit in the European Union's General Data Protection Regulation (GDPR) introduced in 2018. Enforcing some degree of X-AI allows for this and much research has gone into this area of AI (Arrieta et al., 2020). The approach of developing specific algorithms to quantify the degree of influence of *inputs* on *outputs* when supplied with trained black box models (i.e., sensitivity analysis) also holds promise (Datta et al., 2016). However, this is but one tool available to open up black box AI (Guidotti et al., 2018). In Section 3 we describe what cognitive architectures can offer beyond these in applying common sense reasoning and understanding of concepts such as emotions, well-being, and empathy, among others.

3 Cognitive Architectures

In AI, the idea of the mind as a collection of agents (i.e., many cognitive agents or processes / sets of processes that are nearly incomprehensible from one another) is not new, but the diversity in ways to think, represent, and act is still lacking in today's narrow-minded AI. A narrow brute force approach that relies on machine learning instead of strong AI can backfire from an ethical perspective. For example, biased training data can lead to ethical issues such as discrimination. In many instances in life there is no or little training data available (Marcus & Davis, 2019) and this is true also for common sense knowledge. On the other hand, focusing and implementing cognitive architectures can help in dealing with uncertainty, and incomplete and inconsistent information.

While the influence of dual process theory[15] in AI and the social sciences cannot be not understated (Milli et al., 2019) [16], such black and white "dumbbell" thinking tends to constrain thought and theory if taken too literally (Minsky, 1988). There is no reason that we need to stop with just two categories of human cognition; as if all deliberative, reflective, and meta-cognitive (thinking about thinking) cognition is of the same sort and should be classed as such. Stanovich (2004) and Evans (2006) argue that whilst most researchers refer to type 1 as a single system, it is actually a set of (autonomous) systems and processes which each fulfil distinct but related cognitive functions. Glöckner and Witteman (2016) for example categorise intuition (one form of type 1 thinking) into associative, analogous, accumulative, and constructive intuition, discussing how this differentiation helps clarify and provide deeper insight into the relationship between intuition and decision-making. We can also split type 2 thinking into *at least* algorithmic and reflective cognition as Stanovich (2009) does however, we may need to dig deeper if we are to truly understand the full breadth and depth of human behaviour we see out in the real world.

Table 1 below provides one such example of a general cognitive architecture developed by BICA Society (Biological Inspired Cognitive Architectures) with 5 levels of cognition. From 'low' to 'high' level cognitive functions this includes reflexive, reactive/adaptive, proactive/deliberative, reflective, and meta-cognitive/self-aware. This clearly goes beyond the type 1, type 2 dichotomies by separating out cognitive functions which may more or less sit between or across the typical dual-process distinction, allowing for a richer account of complex behaviours such as motivations, emotions, empathy, and more.

[15] For example, Kahneman's (2011) system 1, system 2 dichotomy discussed in *Thinking, Fast and Slow*. Others such as Samuels (2009) propose a type 1, type 2 dichotomy instead which allows to further breakdown the distinctions between and within 'fast' and 'slow' thinking. Marcus and Davis (2019) prefer terms reflexive and deliberative as they are more mnemonic.

[16] Dual process thinking also features heavily (explicitly or not) in theories of moral judgement and reasoning. For example, see Greene (2009, 2014) and Smith (2015). Others contend dual process theories are too crude and that 'higher order' systems must exist to provide meta-control over other cognitive functions, e.g., Sauer (2019) as discussed in *Moral Thinking, Fast and Slow*. Freud (1960) had already three systems in place in his view of the mind: not just ego and id, but also super-ego, the ethical component of the mind.

Table 1: Hierarchy of Cognitive Architectures

Cognitive Function	Purpose	Level
Meta-cognitive and self-aware	Modelling mental states of agents, including own mental states, based on "self" concept.	Highest
Reflective	Modelling internally the environment and behaviour of entities and objects in it.	High
Proactive or Deliberative	Reasoning, planning, exploration, and decision making.	Middle
Reactive or Adaptive	Sub-cognitive forms of learning and adaptation.	Low
Reflexive	Pre-programmed behavioural responses.	Lowest

Cognitive architectures are a subset of agent architectures[17] and come in symbolic (i.e., top-down approach), connectionist (or emergent, i.e., bottom-up approach), or some hybrid combination of these (e.g., serial, or parallel processing, modularised designs, layered hierarchical systems). Over the years, there have been some estimated three hundred cognitive architectures proposed and developed to varying degrees – ranging from those which are solely conceptual through to those which are practically realisable and those actually realised. Of those, Kotserube and Tsotso (2020) survey 84 cognitive architectures developed over the last four decades and clustered them by their perception modality, attentional mechanisms, memory organization, types of learning, and practical applications. Some of the most popular and cited cognitive architectures include ACT-R, Soar, CLARION, ICARUS, EPIC, and LIDA. Others in Duch et al. (2008), Samsonovich (2010), Thórisson & Helgasson (2012), Goertzel et al. (2014), and Ganesha and Venkatamuni (2017) provide additional surveys of the literature on cognitive architectures – a wide, interdisciplinary, and varied body of work. What becomes clear is the lack of general consensus on what cognitive premises and assumptions to work from and further, on any unified theory of cognition in the first place. Future AI – the general sort of AGI envisioned by technology theorists, futurists, and science-fiction writers – requires a broad set of competencies and ways to think, and the ability to regulate behaviour and choose between alternative sets of action possibilities. However, the exact criteria to ascertain AGI is

[17] See Chin et al. (2014) for an overview of the three broad types of agent architectures: *classical*, *cognitive*, and *semantic*. *Classical* including logic, reactive, BDI, and hybrid architectures. *Cognitive* building intelligent agents based on insights from the cognitive sciences (modularised and multi-faceted cognitive designs). *Semantic* combining semantic analysis, NLP, knowledge graphs, and semantic logic.

debated. For example, Newell's (1980, 1992) criteria for AGI includes adaptive behaviour, real-time operability, (normative) rationality, a deep and broad knowledge base, learning, development, linguistic capabilities, self-awareness, and brain realisation. Sun's (2004) desiderata for AGI cognitive architectures in addition includes ecological realism, bio-evolutionary realism, cognitive realism, routinisation, and diversity of methodologies and techniques (and synergistic interactions between them). Adams et al. (2012) suggests perception, memory, attention, reasoning, mobility, planning, motivation, learning, emotion, communication, social interaction, modelling self and others, creativity, and arithmetic criteria. Further, Minsky et al. (2004) propose nine types of reasoning required by future AI: spatial, physical, bodily, visual, psychological, social, reflective, conversational, and educational. The most common (core) set of competencies however includes perception, learning, reasoning, decision-making, planning, and acting (Metzler & Shea, 2011). Others in Vernon et al. (2007), Langley et al. (2009), and Asselman et al. (2015) suggest yet even more AI characteristics to focus on.

Worth acknowledging is that cognitive architectures are complex systems (Menzel & Giurfa, 2001; Schmid et al., 2011) which themselves are generally hierarchical in structure (Simon, 2001) and nearly decomposable (Simon, 1962). In other words, we are able to make distinctions between sub-systems (levels) which themselves are nearly independent, but still intertwined and hence evolve together in time (and space). The higher (slower and coarser) levels in the system conserve stability whilst the lower (faster and finer) levels allow novelty and testing of innovations, mutations, and adaptations to challenges and opportunities in the agents' world. Newell (1990) observed that human activity unfolds on different levels of cognitive processing and can be grouped by timescales at 12 different orders of magnitude (starting at 100μs and up to months/years), providing support to hierarchical thinking and cognition[18]. In the same way that many human systems are complex hierarchical structures, the human mind comprises a set of nested cognitive functions which have evolved over time to help us solve new problems as we face them (Hawkins, 2021). The economy is always discovering, creating, and in process (Arthur, 2006), new problems constantly arise and are solved (or linger). These new problems demand attention and novel thinking to solve them in a never-ending act-react-adapt cycle. Of course, we also leverage our knowledge and past experiences in doing so, highlighting the adaptive and path-dependent nature of intelligence.

[18] Newell (1990) grouped human activity into four bands: biological, cognitive, rational, and social. The highest band, social, includes higher-order abilities such as organisational behaviour, and moral and ethical reasoning.

Cognitive architectures have been deployed in a variety of domains and operational contexts (Kotseruba et al., 2016); human performance modelling, games and puzzles, robotics, psychological experiments, and natural language processing, to name a few. What is lacking from existing cognitive architectures (or at least the current implementations of them) is much of what makes us human (Langley, 2017): the ability to understand and interpret imprecise and complex concepts (e.g., based on common sense, analogous thinking, or abductive reasoning), dynamic memory and continual (online) learning (Diaz-Rodrigues et al., 2018), creativity, emotions and metacognition (and the interactions between them), personality and goal reasoning, and motivation (Dörner & Güss, 2013; Güss & Dörner, 2017). Further, common sense is a rare element for today's AI but is crucial to achieving human-level (general) intelligence (McCarthy, 2007). Common sense requires the sort of (implicit) knowledge that children seem to grasp readily but of which machines struggle greatly. For example, consider the following: *Jack walked into the loungeroom and picked up the TV remote. Jack then walked into the kitchen. Where is the TV remote now?* Despite it not being explicitly stated, one generally assumes the answer is of course *the kitchen*. This example may seem trivial but the mechanisms underlying our understanding and comprehension of the situation are not. Generally speaking, common sense requires many ways to think, and cognitive architectures can offer this diversity in ways to reason about the world around us (Lieto et al., 2018; McCarthy et al., 2002; Singh et al., 2004; Shylaja et al., 2017).

Minsky (2000) stresses the importance of (many) knowledge representations, (relevant) knowledge retrieval, negative expertise (i.e., learning from failures), and self-reflection for common sense AI. Without these we would learn constricted models of the world, not learn from our mistakes, not reflect on what went right or wrong, and even if we did learn, there is no guarantee we could call upon these lessons learned in future analogous problems and situations. Common sense is conditional on a common sense knowledge base; knowledge about the world and things around us that we as humans usually assume are obvious (e.g., grass is green, sky is blue, fire is hot). Such common sense is amassed through experiencing the world and interacting with it. Pearl and Mackenzie (2018) propose a "ladder of causation" (seeing, doing, and imagining[19]) that intelligent agents could use to model how the world works. To be practical, this also requires linking knowledge to uses, goals, or functions (Minsky, 2000). In

[19] Pearl and Mackenzie (2018) contend that most animals (and ML systems) engage only in *seeing* (i.e., associative reasoning). Animals demonstrating a higher cognitive aptitude may engage also in *doing* (i.e., interaction-driven reasoning) that allows evidence-based learning. However, *imagining* (i.e., retrospective reasoning) requires counterfactuals and also requires that *seeing* and *doing* ways to model the world already exist.

other words, defining knowledge by its usefulness to us in getting what we want. For example, augmenting "… knowledge with additional kinds of procedural and heuristic knowledge, such as descriptions of (1) problems that this knowledge item could help solve; (2) ways of thinking that it could participate in; (3) known arguments for and against using it; and (4) ways to adapt it to new contexts" (Minsky et al., 2004). Furthermore, to understand truly 'human' concepts of emotions, love, envy we require much more sophisticated architectures accounting for different sorts of emotions. For example, a reactive layer for primary emotions (e.g., being frightened), a deliberative layer for secondary emotions (e.g., relief), and a meta-management (reflective) layer for tertiary emotions that are typically associated with humans (e.g., love, excitement, anticipation) (Sloman, 2000, 2001).

Hierarchic structures occur frequently in physical, biological, and social systems alike (Simon, 1962). This provides space for cross-fertilisation of knowledge and insights across disciplinary boundaries. For social systems, interactions between elements and the intensity of them are a defining feature. How social infrastructure (e.g., libraries, parks, community centres) interacts with physical (e.g., roads, bridges, telecommunication networks) and environmental (e.g., natural resources, lakes, rivers and wetlands, rainforests) infrastructures is still a nascent area of research (Latham & Layton, 2019) and could be explored by AI. Simon (1962) again provides an important insight: "if the process absorbs free energy the complex system will have a smaller entropy than the elements; if it releases free energy, the opposite will be true" (p. 471). In other words, as long as there is an external source of energy to draw upon, the system will remain relatively stable. However, when energy becomes scarce things may start to unhinge. The social element of human-machine interaction should not be understated; many people anthropomorphise AI agents (Salles et al., 2020) and hence interact with them in ways which require social skills that emulate at least some level of 'human' competency. Acknowledging this then requires a representation of the social processes of behaviour and decision-making that are implementable by algorithms[20]. This also requires understanding the various drivers of human social behaviour such as social norms and status, political viewpoints, and culture, among others. Further, to then be able to identify and correct/adjust for these differences in real-time. Even in the social sense, the way we interact with other people in society resembles a rather well-defined hierarchic structure (Simon, 1962). Operationally,

[20] See for example, Helbing & Molnar's (1994) formulation of *social force* acting on individual agents (of subpopulation *a*) and the interactions between individuals in a subpopulation which simultaneously shape the *social force*. See also Moussaid et al. (2009) for an individual-based model of collective attention; the processes people confronted with information overload use – for better or for worse – to guide them in what to focus on.

"[t]he groupings in this structure may be defined … by some measure of frequency of interaction in this sociometric matrix" (Simon, 1962, p. 469), at least theoretically speaking we should be able to computationally interpret social hierarchy and structure.

Event calculus (Shanahan, 1999; Brandano, 2001) and variants of it (Sadri & Kowalski, 1995; Miller & Shanahan, 2002) provide intuitive way to reason about the world around and how our actions (and the actions of others) may influence it. This also provides a base from which common-sense reasoning can emerge (Mueller, 2014). Essentially, it provides a way to reason about events (e.g., a person crossing the road), fluents (i.e., time-varying property of the world such as the location of a car driving down the road as it approaches the person crossing it), and timepoints (i.e., representing an instant of time like 8.00am on a Monday morning). This offers a very generalisable toolkit to reason about the world, others, and yourself. They can be used in individual and collective accounts as 'narratives' – the horsepower – are a flexible conceptual tool. Tools like event calculus provide a method of reasoning about action and change on a timeline which actual events occur (Mueller, 2008). Situational calculus on the other hand explores hypotheticals and requires more complete specification of hypothetical actions on outcomes (Kowalski & Sadri, 1997). This could allow counterfactual reasoning about the world in imagined scenarios which may or may not ever transpire. Event calculus appears more organic in the sense that allows an incomplete 'narrative' to be specified over perfectly specified situations and demonstrates the clear path-dependency the pervades human life. However, both event and situation calculus support "context-sensitive effects of events, indirect effects, action preconditions, and the common-sense law of inertia" (Mueller, 2008, p. 671), demonstrating their usefulness in AI for human society. For example, the representation and monitoring of social commitments as demonstrated by Chesani et al. (2013). Further, we can use event calculus to represent the goals, values, motivations, and intent of our AI and link them through symbolic nets to events, outcomes, percepts, and actions. These can then be represented to humans in interpretable forms by leveraging the intuitive and descriptive nature of event calculus. This allows better understanding and interpretation by a wider audience. Strangely enough, event calculus is largely absent from mainstream literature on cognitive architectures despite holding promise in areas that plague today's cognitive architectures.

4 Cognitive Architectures for AI Ethics

Cognitive architectures are closely linked to ethics as internal representation deals with aspects such as goals, preferences, desires, and beliefs. It is important to understand how humans and

AI may (complementarily or not) integrate with each other and other cognitive artifacts in their social, cultural, and material environment. After all, ethical and moral reasoning is also heavily influenced by such mechanisms (Awad et al., 2018). Large variations across different social and cultural groups suggests that reasoning about problems with moral and ethical implications is influenced by the presence (or not) of social institutions and cultural norms (e.g., norms of fairness, and collectivism or individualism). Experimental evidence from behavioural economics shows us that cooperation, sharing, and punishment behaviours closely correspond to membership and engagement with cultural groups (Henrich et al., 2001). In other words, we act in similar ways to those we identify with. Despite this, systems thinking for ethical issues rarely goes beyond a single issue or initiative. This is likely because it is very difficult to engage in recursive reasoning about such complex systems, feedbacks, feedforwards, and interactions. For many, it is difficult to conceive how individual actions can relate to ecosystem outcomes and in turn, how ecosystems influence our behaviours. Further, how networks of individuals develop and sustain emergent properties which are not attributed merely to the sum of individual actions (Holling, 2001). Understanding the degree and varieties of cognitive integration between cognitive agents and artifacts requires focusing on dimensions such as information flows, reliability, durability, trust, procedural transparency, informational transparency, individualisation, and transformation (Heersmink, 2015). This can be directly applied to research on distributed morality[21] (Floridi, 2013) and may also advance the study of global collective behaviour[22]. Moral rules and ethical guidelines are distributed throughout groups of like and frequently interacting people (e.g., countries, cultures, families, friend groups, organisations) and some rules propagate widely across groups (e.g., fairness, empathy, cooperation) due to their ability to provide conditions which support peace and prosperity for all. How humans and AI that are engaging in moral and ethical reasoning interact with cognitive artifacts (e.g., social and cultural norms), and the degree by which their decisions are informed by them is crucial to shining light on the black box problem. Marcus and Davis (2019), for example, stress that

> the decisions that the program is making, being computed 'algorithmically,' have an aura of objectivity that impresses bureaucrats and company executives and cows the general public. The workings of the programs are mysterious – the training data is confidential, the program is proprietary, the decision-making process is a "black box" that even the program designers

[21] For example, shared ethics in cultures or local populations to uphold shared values, or moral responsibilities and actions shared between humans and AI agents.
[22] See, for example, Bak-Coleman et al. (2021).

> cannot explain – so it becomes almost impossible for individuals to challenge decisions that they feel are unjust" (p. 36).

Procedural transparency (i.e., transparency in methods, processes, and procedures used to make decisions) is also beneficial in (at least some) public governance and policy settings (Cucciniello et al., 2017) and it is thought to underpin much of democratic society. An intelligent future would benefit from this also and cognitive architectures offer a path forward.

Altman (2014) shows us that mental models shape the choices we make relating to social, economic, and moral problems. Kahneman and Tversky (1984) show us the way problems and choices are framed (i.e., choice architecture) also influences the ordering of preference between alternatives. This begs caution and consideration in how human designers frame ethical goals and values to cognitive architectures and also how they should frame moral problems, situations, action, and events in the first place. The key here is that differences in how we perceive and reason about these problems arise from the mental models we begin with. They act as filters for what things we can perceive, they simplify reasoning and decision making, and they are sensitive to manipulation positive or otherwise. In society, we see how subscribers of certain narratives act in collectively coherent ways that are consistent with whatever cause they subscribe to. There is evidence such narratives play a crucial role in many (socio)economic fluctuations (Shiller, 2017, 2019), in ethics, and in moral reasoning (Roberts, 2012; Brody & Clark, 2014). The problem is when maladaptive ones rise and take hold; these beasts can wreak havoc on peoples' well-being and livelihood and bring destructive force to human social and societal systems, well-intentioned or not. Avoiding negative outcomes (and promoting positive ones) of AI requires keeping a watching eye over potential maladaptive ethical and moral traits forming in a society of humans *and* AI, devising interventions to rectify issues (if any), and if desired, carrying these plans out. This could be supported by a "new[er] kind of AI program—oversight programs—that will monitor, audit, and hold operational AI programs accountable" (Etzioni & Etzioni, 2016). Equally, to hold humans involved in the process accountable also.

As mentioned earlier, cognitive architectures allow to answer *why* questions and put weight on the ability to envision alternative options and realities and compare them. In the next three sections, we explore *why* cognitive architectures are relevant for AI ethics principles: transparency, explainability, and accountability. For each we highlight elements of AI ethics in design, by design, and for design. Further, we discuss how cognitive architectures offer unique advantage and insight compared to other AI technologies.

4.1 Transparency

In relation to cognitive artifacts, Heersmink (2015) suggest two types of transparency that are important for technology (mis)use: *procedural* transparency and *informational* transparency. *Procedural* transparency is the effort and attention (or lack thereof) required to deploy the cognitive artifact. In other words, the cognitive resources required to engage in a certain form of cognitive reasoning (e.g., arithmetic, logic, planning) or ability (e.g., write with pen and paper, drive a car). Experience and learning lend us a hand in transitioning from manual to more autonomous and natural interactions with our cognitive artifacts. *Informational* transparency is the effort and attention (or lack thereof) required to interpret and understand information provided or enabled by the artifact (i.e., its explicability). Together, this transparency enables auditing and monitoring of AI systems' design, use, and implementation. As an artifact for human use, AI will require both forms of transparency for continued adoption. Walmsley (2020) describes two broader varieties of *AI* transparency: *outward* transparency (i.e., concerning the relationship between AI system and things external to it), and *functional* transparency (i.e., concerning the inner workings of the AI itself). In general, Walmsley (2020) notes the current focus on outward transparency as opposed to functional; the greater challenge (technically speaking) of the two. Cognitive architectures are mainly focused on addressing functional transparency but also provide a medium for improving outward transparency by communicating clearly (in succinct, lay terms) goals, functions, and objectives at each layer of an architecture. Further, there are outward and functional elements to consider for both procedural and informational transparency of AI as an artifact for intelligent society. Beyond a purely technical connotation, AI transparency can help address legal aspects of proprietorship, social aspects of interpretation, and user data, algorithmic, and functional literacy, among other issues by looking to the social sciences, law (Larsson & Heintz, 2020), and beyond. This highlights contributions primarily to AI ethics *in* and *by* design. *For* design, emerging standards such as IEEE's P7001 Transparency of Autonomous Systems attempt to fill this gap.

At the heart of AI transparency terminology is opening up the black box of AI, and many X-AI projects (see section 4.2) for how to achieve this technically have been put forward (Adadi & Berrada, 2018; Arrieta et al., 2020) so, *why cognitive architectures?* Cognitive architectures require literally mapping how distinct cognitive functions should work together, ripping apart the black box from within so as to say. By mapping out how AI should work and function as ethical agents we force ourselves to reflect also on our own motivations and the desires of

others as we determine how our AI should think and act. Cognitive architectures force us to commit to certain cognitive assumptions, presumptions, processes, and artifacts which then provide basis for auditability if these are made explicit and documented. This also provides testable theories and hypotheses. The commitment (at each and every level of cognition) to documented goals, objectives, functions, and assumptions support ethics *for* design and auditability. Ethics *in* design is addressed by providing a communicative medium by which engineers, policymakers, and end-users (e.g., the public) have a common language and means for interpretation, deliberation, and engagement. This also contributes the tried-and-true AI tools and methodologies of the cognitive architecture literature. Finally, ethics *by* design is advanced by cognitive architectures which support a wider, more general and human-like cognitive basis for artificial and autonomous moral agents.

4.2 Explainability

Explainability is a neighbouring concept to transparency albeit with narrower scope and more technical focus (Larsson and Keintz, 2020), and is typically defined on a model, component, or algorithmic scale (Lepri et al., 2018). It is concerned with the concepts and methods required for human interpretation of AI systems and requires that when queried, the AI "… must be able to retrieve the ways in which its choices relate to norms and then communicate them in accessible terms" (Langley, 2019, p. 9778). What is considered adequate and accessible will vary for different users and situations. For example, the engineer who is tasked with fixing a faulty or unethical AI will require different forms of communication and explanation (e.g., to debug) as compared to an end-user (e.g., member of the public) who may simply want to understand how inputs *relative to them* turn into outputs again *relative to them*. The engineer may already have requirements to meet at the model-level provided to them (e.g., reduce likelihood of algorithmic discrimination to minorities to some pre-determined level deemed 'acceptable' by the company) and so, will turn focus to the component-level first to figure out which module of the AI may be malfunctioning or introducing bias, and then zoom in to the algorithmic-level of the troubled module(s). In contrast, the end-user may be satisfied with a model-level explanation with less need for technical details. Such personalisation of AI requires understanding and consideration of elements such as culture and emotion which cognitive architectures show promise (Sun, 2020; Samsonovich, 2020) – an AI ethics *by* design contribution.

When we ask AI to explain itself, we (implicitly or not) assume that AI has reasons for what it does. Humans do (Searle & Willis, 1983), so why wouldn't AI? Without delving too far down

this rabbit hole, some researchers have questioned whether AI will ever truly understand what it is processing or thinking about and whether it can demonstrate actual intentionality. The degree to which AI understand and demonstrate intentionality can be extended by focusing on cognitive architectures (Chong et al., 2007). For example, the ACT-R architecture (Anderson et al., 2004) includes a distinct intentional module to carry out goals from desires and Icarus (Langley and Choi, 2006) casts intentions as situation-specific instances of more general long-term memory and knowledge structures. Given the modern powers of AI and recent strides in deep learning (alongside enabling and emerging information technologies[23]), it is expected the "trade-off between efficiency and thoroughness will [continue to] move far toward thoroughness" (Hickey, 2016, p. 94). If thoroughness in modelling and understanding the human mind is our aim, cognitive architectures often explicitly state their aim is to model or emulate human behaviour. They also offer an intuitive interpretation by visual systems-level representations and descriptive code (e.g., event calculus) – an ethics *for* design contribution. Goals, values, and intentions can be explicitly designed and specified and hence, the degree to which AI achieves intentionality and understanding is only really limited by the complexity of our architectures, limitations of computational resources, and creativity on part of the designer. X-AI is "not only invested in how to structure the mind, but how we as agents understand other minds to work" (Westberg et al., 2019, p. 210) lending support also to self-reflective and meta-cognitive processes in AI cognitive architectures. In other words, the architectures to allow how to think about others' thinking and further still, deciding what then to do with these beliefs and expectations about others' thinking and acting on those decisions. As we build systems which can reflect and consider from others' viewpoints, we may also come to reflect on our own thinking and potential biases as Richards (2019) has suggested – contributing to ethics *in* design.

4.3 Accountability

Accountability requires transparency (among other things[24]) on the AI design and stakeholder engagement and deliberation process for a credible and clear assignment of responsibility and accountability throughout the AI lifecycle. This is particularly true in legal situations where mishaps and incidents occur which cause harm to human life and livelihood and where

[23] For example, quantum technologies for AI and models of the mind (Bickley et al., 2021).

[24] Fox (2007) provides evidence to support rejecting the assumption that transparency generates accountability under all circumstances. Transparency is found to be "necessary but far from sufficient to produce accountability" and we should instead seek answers to questions like "under what conditions can transparency lead to accountability" and "what types of transparency manage to generate what types of accountability" (p. 665).

punishments and compensations may need to be made. However, this depends on our motivations for accountability. For example, do we focus on the individual(s) at fault (e.g., lower-level workers and operators) or focus on the institution(s) as a whole (e.g., company, professional body, watchdog, government department). In the human error literature[25], this coincides with the person- and systems-level approaches to human factors, as introduced by Reason (2000). The person approach focuses on processes, errors, and biases at the *individual* level and assigns blame to errors made by those closest to the proximal cause of the incident under investigation (i.e., bottom-up approach). In contrast, the systems approach focuses on evidence of systemic weaknesses in the institution itself that may have contributed to the incident either actively or latently (i.e., top-down approach)[26]. Integrating these two approaches (person and systems) requires zooming in and out by (1) *outward transparency* for understanding how design and implementation decisions on the part of humans have contributed to hazardous situations, and (2) *functional transparency* for understanding the AI contributions (and more finely grained split of designer/coder/AI contributions as well). To also promote the beneficial and advantageous outcomes for business and society we should still look to protect intellectual property or rights to privacy where this may be of concern (e.g., risks of gaming the system exist, proprietary information, sensitive data). Accountability in these situations may necessitate a more opaque transparency of data, algorithms, or models at certain cognitive levels. Counterfactuals show promise to provide information about how changes to inputs affect outputs as this provides contestable basis without necessarily requiring complete transparency (Doshi-Velez et al., 2019; Wachter et al., 2017). Instead, they can show how subtle changes to inputs, context, and environment can lead to drastic changes in outcomes via network effects and feedback loops between AI designers, researchers, builders, regulators, and society. The process of designing a cognitive architecture (if documented) itself generates the evidence required to know which risks were considered, what the goals and objectives of design were, how the relevant stakeholders were identified and engaged, and what cognitive assumptions and biases are built into the AI from the get-go.

[25] A subset of the human factors and ergonomics (HFE) literature focused on human and organisational inputs to hazardous and unsafe outcomes in human sociotechnical systems. It is popular in high-risk, high-reliability domains such as aerospace, medicine, maritime, mining, rail, and heavy industry.

[26] Gasser and Almeida (2017) provide a good start with their layered approach (social & legal, ethical, technical) to the AI governance system with indicative temporal scale (near-, medium-, and long-term timing).

5 Concluding Remarks and Future Perspectives

In this paper we have emphasized that an AI that is ethical and reliable will require strong and general AI that considers cognitive architecture aspects in order to function in our complex and constantly changing world with so many unpredictable elements that are hard to program and anticipate. In recent years we have obtained wonderful progress with an AI that is narrow which distracts for the importance of a cognitive architecture as a way of providing AI with a deeper understanding of the world and those in it. The great success in well-defined areas (e.g., board games such as chess or Go or game shows such as Jeopardy!) cannot be scaled up to a complex real-world environment. We also need to get better (theoretical) understanding why Deep Learning works so effectively in those specialized circumstances. For example, more progress will be achieved in understanding the power and characterization of multi-layer neural networks and how to reason about meta-level reasoning (for a discussion, see, e.g., Perez, 2018). We therefore need to be open to a large set of tools and methods to deal with such complexity. Marcus and Davis (2019) point out that "what we have for now are basically digital idiots savants" (p. 13) criticizing that "we ceding more and more authority to machines that are unreliable and, worse, lack any comprehension of human values. The bitter truth is that for now the vast majority of dollars invested in AI are going towards solutions that are brittle, cryptic, and too unreliable to be used in high-stakes problems" (p. 15). They argue we are experiencing a short-term obsession with narrow AI that goes for the low-hanging fruits rather than the more challenging and long-standing problems. A focus on cognitive architectures can counteract such tendencies. It may also provide ways of thinking how cognitive architectures can be improved by insights from approaches such as deep learning. For example, cognitive architectural approaches have struggled historically in incorporating learning elements. But cognitive architectures are essential as the real world is an open system that requires constant adjustments to what is changing in its surroundings. Future research could explore in more detail what we can learn in the area of open and emergent systems and complexity research as, for example, done by scholars at the Santa Fe Institute (see, e.g., Krakauer, 2019) to transfer those insights into AI. This could help AI to better tackle the world in all its complexity and richness.

We are still in the process of understanding how to analyse evolving or ever-unfolding systems, systems in which we do not know what can and will happen next which puts a toll on our current tools of thought and exploration that strongly emphasize the virtue of rationality and reason. AI capable of reasoning about situations with moral and ethical elements require

common sense to allow adaptation to context-specific issues and considerations and hence, need to remain responsive to the environment and those in it. For example, for AI to abide by even Asimov's first law[27] requires it first understand the meaning of causing harm to humans. This requires may ways of thinking and also representing knowledge on many layers and abstractions. Cognitive architectures which go beyond shallower models of cognition (e.g., system 1, system 2) allow this flexibility and in so, robustness and resilience in facing new problems and situations including those we may not have ever experienced before (i.e., learning of the fly, 'winging' it). Adding common sense reasoning into AI systems will help to reduce system vulnerabilities or sabotage attempts that have substantial ethical implications. In addition, common sense reasoning is a step towards thinking how social or emotional intelligence can be included. But for that, we also need to improve our computational theories around psychological processes. The *Moral Competence in Computational Architectures for Robots* initiative financed by the Department of Defence including scholars from Tufts University, Brown University, RPI, Georgetown University, and Yale University, for example, tries to develop computational architectures that are capable of moral reasoning via identifying the "logical, cognitive, and social underpinnings of human moral competence"[28].

No doubt there will be shortcomings and pitfalls of AI. This is to be expected as it has occurred time again with other foundational technologies in history. Especially those in the earlier stages of mass technological adoption. Crucial for wider adoption and integration in human life and society is transparency, explainability, and accountability for AI in design right through to implementation and upkeep (considering the full lifecycle). These require opening the black box of AI, something we have argued that cognitive architectures show promise. This requires to also go beyond the power of data and focus on aspects such as the ethical values that designers and programmers use to make a fair, transparent, and safe world. Also, the means by which we communicate and deliberate these to a wider audience of technical and non-technical stakeholders alike.

[27] *A robot may not injure a human being or, through inaction, allow a human being to come to harm.*
[28] https://hrilab.tufts.edu/muri13/

Declarations

Ethics approval	Not applicable.
Consent to participate	Not applicable.
Consent for publication	Not applicable.
Availability of data and materials	Any data and materials used in the study are available on request to the corresponding author.
Code availability	Any code used in the study are available on request to the corresponding author.
Conflicts of interest	The author(s) declare no conflicts of interest or competing interests.
Funding	This research is/was supported by an Australian Research Training Program (RTP) Scholarship.

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access*, *6*, 52138-52160.

Adams, S., Arel, I., Bach, J., Coop, R., Furlan, R., Goertzel, B., Hall, S., Samsonovich, A., Sheutz, M., Schlesinger, M., Shapiro, S., & Sowa, J. (2012). Mapping the landscape of human-level artificial general intelligence. *AI magazine*, *33*(1), 25-42.

Altman, M. (2014). Mental models, decision-making, bargaining power, and institutional change. World Interdisciplinary Network for Institutional Research Conference, Greenwich, London, UK. https://winir.org/?page=conferences&side=winir_2014.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, *111*(4), 1036.

Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82-115.

Arthur, W.B. (2006). The Nature of Technology: What It Is and How It Evolves. Free Press, New York, US.

Asselman, A., Aammou, S., & Nasseh, A. E. (2015). Comparative study of cognitive architectures. *International Research Journal of Computer Science (IRJCS)*, *2*(9), 8-13.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, F., & Rahwan, I. (2018). The moral machine experiment. *Nature*, *563*(7729), 59-64.

Bak-Coleman, J. B., Alfano, M., Barfuss, W., Bergstrom, C. T., Centeno, M. A., Couzin, I. D., ... & Weber, E. U. (2021). Stewardship of global collective behavior. *Proceedings of the National Academy of Sciences*, *118*(27), e2025764118.

Bench-Capon, T., & Modgil, S. (2017). Norms and value based reasoning: justifying compliance and violation. *Artificial Intelligence and Law*, *25*(1), 29-64.

Berendt, B. (2019). AI for the Common Good?! Pitfalls, challenges, and ethics pen-testing. *Journal of Behavioral Robotics*, *10*(1), 44-65.

Bickley, S.J., Chan, H.F., Schmidt, S.L., & Torgler, B. (2021). Quantum-sapiens: the quantum bases for human expertise, knowledge, and problem-solving. *Technology Analysis & Strategic Management*, 1-13. DOI: 10.1080/09537325.2021.1921137.

Brady, N., & Hart, D. (2007). An exploration into the developmental psychology of ethical theory with implications for business practice and pedagogy. *Journal of Business Ethics*, *76*(4), 397-412.

Brandano, S. (2001). The event calculus assessed. In *Proceedings Eighth International Symposium on Temporal Representation and Reasoning (TIME 2001)*, 7-12. IEEE Computer Society, Washington, US.

Brody, H., & Clark, M. (2014). Narrative Ethics: A Narrative. *The Hastings Center Report*, *44*(s1), S7–S11. DOI: 10.1002/hast.261

Bryson, J.J. (2020). The Artificial Intelligence of the Ethics of Artificial Intelligence. In M.D. Dubber, F. Pasquale & S. Das (Eds.) *The Oxford Handbook of Ethics of AI* (p. 1). Oxford University Press, UK.

Carabantes, M. (2020). Black-box artificial intelligence: an epistemological and critical analysis. *AI and Society*, *35*(2), 309-317.

Cervantes, S., López, S., & Cervantes, J. A. (2020). Toward ethical cognitive architectures for the development of artificial moral agents. *Cognitive Systems Research*, *64*, 117-125.

Chen, J., & Barnes, M. (2014). Human-agent teaming for multirobot control: a review of human factors issues. *IEEE Transactions on Human-Machine Systems*, *44*(1), 13-29.

Chesani, F., Mello, P., Montali, M., & Torroni, P. (2013). Representing and monitoring social commitments using the event calculus. *Autonomous Agents and Multi-Agent Systems*, *27*(1), 85-130.

Chong, H.Q., Tan, A.H., & Ng, G.W. (2007). Integrated cognitive architectures: a survey. *Artificial Intelligence Review*, *28*(2), 103-130.

Chrisley, R. (2020). "A Human-Centered Approach to AI Ethics." In M.D. Dubber, F. Pasquale, S. Das (Eds.) *The Oxford Handbook of Ethics of AI* (p. 463-474). Oxford University Press, UK.

Collingridge, D. (1980). The social control of technology. Frances Pinter, London.

Costello, T., & McCarthy, J. (1999). Useful counterfactuals. *Linköping Electronic Articles in Computer and Information Science*, *3*, 1-28.

Cucciniello, M., Porumbescu, G. A., & Grimmelikhuijsen, S. (2017). 25 years of transparency research: Evidence and future directions. *Public Administration Review*, *77*(1), 32-44.

Datta, A., Sen, S., & Zick, Y. (2016, May). Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *2016 IEEE symposium on security and privacy (SP)*, 598-617. IEEE, California, US.

Díaz-Rodríguez, N., Lomonaco, V., Filliat, D., & Maltoni, D. (2018). Don't forget, there is more than forgetting: new metrics for Continual Learning. *arXiv preprint* arXiv:1810.13166.

Dignum, V. (2019). Responsible artificial intelligence: how to develop and use AI in a responsible way. Springer Nature, Basingstoke.

Dörner, D., & Güss, C. D. (2013). PSI: A computational architecture of cognition, motivation, and emotion. *Review of General Psychology*, *17*(3), 297-317.

Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S., O'Brien, D., Scott, K., Shieber, S., Waldo, J., Weinberger, D., Weller, A., & Wood, A. (2019). Accountability of AI under the law: The role of explanation. arXiv preprint arXiv:1711.01134, version 3.0.

Duch, W., Oentaryo, R. J., & Pasquier, M. (2008). Cognitive architectures: Where do we go from here? *AGI*, *171*(1), 122-136.

Dyrkolbotn, S., Pedersen, T., & Slavkovik, M. (2018). On the distinction between implicit and explicit ethical agency. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (p. 74-80), New Orleans, USA.

Etzioni, A., & Etzioni, O. (2016). AI assisted ethics. *Ethics and Information Technology*, *18*(2), 149-156.

Evans, J. (2006). Dual system theories of cognition: Some issues. *Proceedings of the 28th Annual Meeting of the Cognitive Science Society* (p. 202-207), Vancouver, Canada.

Fishkin, J.S. (2011). When the People Speak: Deliberative Democracy and Public Consultation. Oxford University Press, UK.

Floridi, L. (2013). Distributed morality in an information society. *Science and Engineering Ethics*, *19*(3), 727-743.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—an ethical framework

for a good AI society: opportunities, risks, principles, and recommendations. *Minds and Machines*, *28*(4), 689-707.

Fox, J. (2007). The uncertain relationship between transparency and accountability. *Development in Practice*, *17*(4-5), 663-671.

Freud, S. (1960). *The Ego and the Id*. W. W. Norton & Company.

Ganesha, D., & Venkatamuni, V. M. (2017). Review on Cognitive Architectures. *Indian Journal of Science and Technology*, *10*(1), 1-8.

Gasser, U., & Almeida, V. A. (2017). A layered model for AI governance. *IEEE Internet Computing*, *21*(6), 58-62.

Glöckner, A., & Witteman, C. (2010). Beyond dual-process models: A categorisation of processes underlying intuitive judgement and decision making. *Thinking & Reasoning*, *16*(1), 1-25.

Goertzel, B., Pennachin, C., & Geisweiller, N. (2014). Brief survey of cognitive architectures. In *Engineering General Intelligence*, *Part 1* (p. 101-142). Atlantis Press, Paris.

Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, *38*(3), 50-57.

Greene, J.D. (2009). The cognitive neuroscience of moral judgment. In M.S. Gazzaniga, G.R. Mangun (Eds.), *The Cognitive Neurosciences*, 1013–1023. MIT Press, US.

Greene, J.D. (2014). Beyond point-and-shoot morality: Why cognitive (neuro) science matters for ethics. *Ethics*, *124*(4), 695-726.

Griffiths, M. R., & Lucas, J. R. (2016). Value Economics: The Ethical Implications of Value for New Economic Thinking. Springer, New York.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR)*, *51*(5), 1-42.

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G. Z. (2019). XAI—Explainable artificial intelligence. *Science Robotics*, *4*(37), eaay7120.

Güss, C. D., & Dörner, D. (2017). The importance of motivation and emotion for explaining human cognition. *Behavioral Brain Sciences*, *40*, 38-39.

Hawkins, J. (2021). A Thousand Brains: A New Theory of Intelligence. Basic Books, New York, US.

Heersmink, R. (2015). Dimensions of integration in embedded and extended cognitive systems. *Phenomenology and the Cognitive Sciences*, *14*(3), 577-598.

Helbing, D. (2019). Machine Intelligence: Blessing or Curse? It Depends on Us! In Towards Digital Enlightenment (pp. 25-39). Springer, Cham.

Helbing, D., & Molnar, P. (1995). Social force model for pedestrian dynamics. *Physical Review E*, *51*(5), 4282.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., & McElreath, R. (2001). In search of homo economicus: behavioral experiments in 15 small-scale societies. *American Economic Review*, *91*(2), 73-78.

Hickey, T.J. (2016). Twentieth-Century Philosophy of Science: A History. T.J. Hickey Publishing, Unknown location.

Holling, C.S. (2001). Understanding the complexity of economic, ecological, and social systems. *Ecosystems*, *4*(5), 390-405.

Homer (1945). The Odyssey. Translation by A. T. Murray. Harvard University Press.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389-399.

Kahneman, D. (2011). Thinking, fast and slow. Macmillan, New York, US.

Kahneman, D., & Tversky, A. (1984). Choices, values, and frames. *The American Psychologist*, *39*(4), 341–350. DOI: 10.1037/0003-066X.39.4.341.

Kotseruba, I., Gonzalez, O. J. A., & Tsotsos, J. K. (2016). A review of 40 years of cognitive architecture research: Focus on perception, attention, learning and applications. arXiv preprint arXiv:1610.08602, 1-74.

Kotseruba, I., & Tsotsos, J. K. (2020). 40 years of cognitive architectures: core cognitive abilities and practical applications. *Artificial Intelligence Review*, *53*(1), 17-94.

Kowalski, R., & Sadri, F. (1997). Reconciling the event calculus with the situation calculus. *The Journal of Logic Programming*, *31*(1-3), 39-58.

Krakauer, D. C. (Ed.) (2019). *World Hidden in Plain Sight: The Evolving Idea of Complexity at the Santa Fe Institute*. SFI Press.

Langley, P. (2017). Progress and challenges in research on cognitive architectures. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)*, 4870-4876. AAAI, California, US.

Langley, P. (2019). Explainable, normative, and justified agency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(1), 9775-9779.

Langley, P., & Choi, D. (2006). A unified cognitive architecture for physical agents. In *Proceedings of the National Conference on Artificial Intelligence*, *21*(2), 1469). Association for the Advancement of Artificial Intelligence (AAAI), Boston, US.

Langley, P., Laird, J. E., & Rogers, S. (2009). Cognitive architectures: Research issues and challenges. *Cognitive Systems Research*, *10*(2), 141-160.

Larsson, S., & Heintz, F. (2020). Transparency in artificial intelligence. *Internet Policy Review*, *9*(2), 1-16.

Latham, A., & Layton, J. (2019). Social infrastructure and the public life of cities: Studying urban sociality and public spaces. *Geography Compass*, *13*(7), e12444.

Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2018). Fair, transparent, and accountable algorithmic decision-making processes. *Philosophy & Technology*, *31*(4), 611-627.

Lieto, A., Lebiere, C., & Oltramari, A. (2018). The knowledge level in cognitive architectures: Current limitations and possible developments. *Cognitive Systems Research*, *48*, 39-55.

Malle, B. F., & Knobe, J. (1997). The folk concept of intentionality. *Journal of Experimental Social Psychology*, *33*(2), 101-121.

Marcus, G. & Davis, E. (2019). *Rebooting AI: Building Artificial Intelligence We Can Trust*. Pantheon Books.

McCarthy, J. (2007). From here to human-level AI. *Artificial Intelligence*, *171*(18), 1174-1182.

McCarthy, J., Minsky, M., Sloman, A., & Gong, L. (2002). An architecture of diversity for commonsense reasoning. *IBM Systems Journal*, *41*(3), 530.

Menzel, R., & Giurfa, M. (2001). Cognitive architecture of a mini-brain: the honeybee. *Trends in cognitive sciences,* 5(2), 62-71.

Metzler, T., & Shea, K. (2011). Taxonomy of cognitive functions. In DS 68-7: *Proceedings of the 18th International Conference on Engineering Design (ICED 11), Impacting Society through Engineering Design*, 7(1), 330-341.Lyngby/Copenhagen, Denmark.

Miller, R., & Shanahan, M. (2002). Some alternative formulations of the event calculus. In A.C. Kakas, F. Sadri (Eds.), *Computational logic: logic programming and beyond*, 452-490. Springer, Berlin, DE.

Milli, S., Lieder, F., & Griffiths, T. L. (2019). A rational reinterpretation of dual-process theories. *Preprint Article*. DOI: 10.13140/RG.2.2.14956.46722/1.

Minsky, M. (1988). Society of mind. Simon and Schuster, New York, US.

Minsky, M. (2000). Commonsense-based interfaces. *Communications of the ACM*, *43*(8), 66-73.

Minsky, M., Singh, P., & Sloman, A. (2004). The St. Thomas common sense symposium: designing architectures for human-level intelligence. *Ai Magazine*, *25*(2), 113-113.

Mitchell, M. (2019). Artificial intelligence: A guide for thinking humans. Penguin UK.

Moor, J. (2006). The Nature, Importance, and Difficulty of Machine Ethics. *IEEE Intelligent Systems*, *21*(4), 18-21.

Moussaid, M., Helbing, D., & Theraulaz, G. (2009). An individual-based model of collective attention. arXiv preprint arXiv:0909.2757.

Mozelewski, T.G., & Scheller, R.M. (2021). Forecasting for intended consequences. *Conservation Science and Practice*, *3*(4), e370.

Mueller, E.T. (2008). Event calculus. *Foundations of Artificial Intelligence*, *3*, 671-708.

Mueller, E.T. (2014). Commonsense reasoning: an event calculus based approach. Morgan Kaufmann, Massachusetts, US.

Müller, V. C. (2021). Is it time for robot rights? Moral status in artificial entities. *Ethics and Information Technology*, 1-9. DOI: 10.1007/s10676-021-09596-w.

Newell, A. (1980). Physical symbol systems. *Cognitive Science*, *4*(2), 135-183.

Newell, A. (1990). Unified Theories of Cognition. Cambridge, Harvard University Press, Massachusetts, US

Newell, A. (1992). Précis of unified theories of cognition. *Behavioral and Brain Sciences*, *15*(3), 425-437.

Owen, R., Macnaghten, P., & Stilgoe, J. (2012). Responsible research and innovation: From science in society to science for society, with society. *Science and Public Policy*, *39*(6), 751-760.

Pariser, E. (2011). The filter bubble: What the Internet is hiding from you. Penguin UK.

Pearl, J., & Mackenzie, D. (2018). The book of why: the new science of cause and effect. Basic books, UK.

Perez, C. E. (2018). Artificial Intuition: The Improbable Deep Learning Revolution. Self-published by C. E. Perez.

Reason, J., (2000). Human error: models and management. *British Medical Journal*, *320*(7237), 768-770. DOI: 10.1136/bmj.320.7237.768.

Richards, D. (2019). Explainable AI: Transparent Pedagogical Agents that help the Learner to Reflect. Unpublished manuscript.

Roberts, R. (2012). *Narrative Ethics. Philosophy Compass*, *7*(3), 174–182. DOI: 10.1111/j.1747-9991.2011.00472.x.

Sadri, F., & Kowalski, R. A. (1995). Variants of the Event Calculus. In *Proceedings of the Twelfth International Conference on Logic Programming (ICLP-95)*, 67-81. Association for Logic Programming, Tokyo, JP.

Salles, A., Evers, K., & Farisco, M. (2020). Anthropomorphism in AI. *AJOB Neuroscience*, *11*(2), 88-95.

Samuels, R. (2009). The magical number two, plus or minus: Dual-process theory as a theory of cognitive kinds. In J. Evans, K. Frankish (Eds.) *In two minds: Dual processes and beyond*, 129-146. Oxford University Press, UK.

Samsonovich, A.V. (2010). Toward a unified catalog of implemented cognitive architectures. *BICA*, *221*(2010), 195-244.

Samsonovich, A.V. (2020). Socially emotional brain-inspired cognitive architecture framework for artificial intelligence. *Cognitive Systems Research*, *60*, 57-76.

Sauer, H. (2019). Moral thinking, fast and slow. Routledge, New York, US.

Schmid, U., Ragni, M., Gonzalez, C., & Funke, J. (2011). The challenge of complexity for cognitive systems. *Cognitive Systems Research*, *12*, 211-218.

Searle, J.R., & Willis, S. (1983). Intentionality: An essay in the philosophy of mind. Cambridge University Press, UK.

Shanahan, M. (1999). The event calculus explained. In M.J. Wooldridge, M. Veloso (Eds.), *Artificial Intelligence Today: Recent Trends and Developments*, 409-430. Springer, Berlin, DE.

Shiller, R.J. (2017). Narrative Economics. *American Economic Review*, *107*(4), 967-1004.

Shiller, R.J. (2019). Narrative economics: How stories go viral and drive major economic events. Princeton University Press, New Jersey, US.

Shneiderman, B. (2020). Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy Human-Centered AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *10*(4), 1-31.

Shylaja, K. R., Vijayakumar, M. V., Prasad, E. V., & Davis, D. N. (2017). Artificial Minds with Consciousness and Common sense Aspects. *International Journal of Agent Technologies and Systems (IJATS)*, *9*(1), 20-42.

Siau, K., & Wang, W. (2020). Artificial intelligence (AI) ethics: ethics of AI and ethical AI. *Journal of Database Management (JDM)*, *31*(2), 74-87.

Simon, H. (1962). The Architecture of Complexity. *Proceedings of the American Philosophical Society*, *106*(6), 467-482.

Simon, H. (2001). Complex Systems: The Interplay of Organisations and Markets in Contemporary Society. *Computational & Mathematical Organisation Theory*, *7*(1), 79-85.

Singh, P., Minsky, M., & Eslick, I. (2004). Computing commonsense. *BT Technology Journal*, *22*(4), 201-210.

Sloman, A. (2000). Architectural requirements for human-like agents both natural and artificial: What sorts of machines can love? In K. Dautenhahn (Ed.), *Human Cognition and Social Agent Technology*. John Benjamins Publishing Company, Amsterdam, NL.

Sloman, A. (2001). Beyond shallow models of emotion. *Cognitive Processing*, *2*(1), 177-198.

Smith, H. (2015). Dual-process theory and moral responsibility. In R. Clarke, M. McKenna, A.M. Smith (Eds), *The Nature of Moral Responsibility: New Essays*, 175-209. Oxford University Press, US.

Stanovich, K.E. (2004). The robot's rebellion: Finding meaning in the age of Darwin. Chicago: University of Chicago Press.

Stanovich, K.E. (2009). "Distinguishing the reflective, algorithmic, and autonomous minds: Is it time for a tri-process theory?" In J. Evans & K. Frankish (Eds.) In two minds: Dual processes and beyond (p. 55–88). Oxford University Press, UK.

Sun, R. (2004). Desiderata for cognitive architectures. *Philosophical Psychology*, *17*(3), 341-373.

Sun, R. (2020). Exploring culture from the standpoint of a cognitive architecture. *Philosophical Psychology*, *33*(2), 155-180.

Thaler, R. H., & Sunstein, C. R. (2009). Nudge: improving decisions about health, wealth and happiness . Penguin, New York, US.

Thórisson, K., & Helgasson, H. (2012). Cognitive architectures and autonomy: A comparative review. *Journal of Artificial General Intelligence*, *3*(2), 1-30.

Van Berkel, N., Tag, B., Goncalves, J., & Hosio, S. (2020). Human-centred artificial intelligence: a contextual morality perspective. *Behaviour & Information Technology*, 1-17. DOI: 10.1080/0144929X.2020.1818828.

van den Hoven, J., Vermaas, P. E., & van de Poel, I. (2015). Design for values: An introduction. In van den Hoven, Vermaas, & van de Poel (Eds.) *Handbook of ethics, values, and technological design: Sources, theory, values and application domains* (p. 1-7). Springer, Dordrecht.

van de Poel, I. (2020). Embedding values in artificial intelligence (AI) systems. *Minds and Machines*, *30*(3), 385-409.

Vernon, D., Metta, G., & Sandini, G. (2007). A survey of artificial cognitive systems: Implications for the autonomous development of mental capabilities in computational agents. *IEEE transactions on evolutionary computation*, *11*(2), 151-180.

Visser, E.J., Pak, R., & Shaw, T.H. (2018). From 'automation' to 'autonomy': the importance of trust repair in human-machine collaboration. *Ergonomics*, *61*(10), 1409-27.

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, *31*, 841-887.

Wallach, W., & Allen, C. (2008). Moral machines: Teaching robots right from wrong. Oxford University Press, UK.

Walmsley, J. (2020). Artificial intelligence and the value of transparency. *AI & Society*, *36*, 585-595.

Westberg, M., Zelvelder, A., & Najjar, A. (2019). A historical perspective on cognitive science and its influence on XAI research. In R. Goebel, Y. Tanaka, & W. Wahlster (Eds.), *International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems* (p. 205-219). Springer, Cham.