Can AI help make us better humans?

Exploring AI for enhanced moral education in early education

Melissa Novitsky

Supervised by Dr. Anna Puzio & Dr. Peter Stegmaier

Master Thesis - 23 September 2024

Master of Science in Philosophy of Science, Technology, and Society

Faculty of Behavioural, Management, and Social Sciences — University of Twente

Enschede, The Netherlands

"*Oppression involves a failure of the imagination: the failure to* ***imagine the full humanity of other human beings.***"

– Margaret Atwood

"Don't stop fightin' and don't stop believin'. You can make the world better for your kids before you leave it. [...]
Corruption always leads us to the same [stuff] again, so when you talk 'bout revolution, dawg, I hear just what you sayin'. What good is takin' over, when we know what you gon' do?
The only real revolution happens right inside of you."

– J. Cole, *'High For Hours'*

-Contents-

Abstract

This master's thesis investigates the integration of responsibly designed technology to enhance moral education within early education practices, with a particular focus on AI and social robots. The central research question explores how these technologies can be utilized to foster moral development, critical self-reflection, and emotional regulation in young learners. The thesis posits that a blend of best practices from a variety of disciplines is essential for a comprehensive educational approach and responsible technological design. It argues that technologies, when designed with ethical principles through an adapted Value Sensitive Design (VSD) approach, can play a significant role in moral education.

The research categorizes the investigation into three main elements: conceptual exploration of early and moral education, technical examination of existing AI implementations, and empirical application of the VSD approach. The conceptual analysis addresses key cognitive development factors in early education and identifies gaps in current moral education practices, particularly the lack of focus on technomoral resilience. The technical investigation includes a case study of the PopBots system, a toolkit for children to engage in AI activities, and reviews recent recommendations for AI's role in moral enhancement. Empirical investigations apply the VSD approach to AI and social robots, tackling ethical challenges such as the black box problem and highlighting opportunities for AI for Social Good (AI4SG). The analysis synthesizes VSD evaluations of social robots in early education, identifying stakeholder values and proposing responsible implementation solutions.

These investigations culminate in the proposed 'MiruBots' project, which combines the effective elements from the technical investigation of the PopBots system and AI moral enhancement strategies to create a practical, enhanced solution for moral education in early education. This project is evaluated against the values identified in both the conceptual and empirical investigations.

The thesis concludes by recapitulating key findings, underscoring the importance of responsible technological implementation, and advocating for further research in this area. By synthesizing separate research domains—moral education using AI, AI ethics in early education, and value concerns of social robots—this thesis aims to present a solution for how to foster critical reflection, open-mindedness, and forethought in educational and innovation development processes, ultimately contributing to a more peaceful and beneficial global society.

Keywords: moral education, early education, AI, social robots, technomoral resilience, VSD

1 Introduction

For anyone who has ever pondered how and why humanity has reached its current state, with its remarkable advancements alongside the ongoing presence of unimaginable atrocities, the thought of how humanity could be improved may have also crossed their mind. One such thought might have been that the global society may be better off further embracing things more aligned with the 'soft' sciences, instead of the prosaic overemphasis of focusing efforts largely on things like the 'hard' sciences (Shapin, 2022). Or better yet, a thought might have been that the commonplace demarcation of many things, like the hard and soft sciences, serves little purpose, and "that the most valued products and practices of late modernity are hybrids" of seemingly opposing natures, of embodying both the technical and the uniquely human (Shapin, 2022, p. 327). A more harmonious and prosperous future for humanity might involve comprehensively embracing the good in both sides of these 'opposing' concepts, with more of an emphasis on a heightened capacity for critical self-reflection and things like compassion and empathy. I set out to investigate how we may use the best of existing technologies in a thoughtful manner which helps to create a more open-minded and good willed society.

Thus, the main research question is as follows: **How could technology be responsibly designed to aid in an enhanced moral education within effective early education practices?**

The method of investigation into this notion will be rooted in the Value Sensitive Design (VSD) methodology, primarily consisting of synthesis and analysis of the available research and studies which encompass the predominant elements of concern within the research question. The conceptual, technical, and empirical investigations of this inquest will result in a suggested solution based largely on an actual case study of significant relevance, enhanced by the findings of recently conducted VSD studies of a highly related nature. This collective methodology, once enacted, yields answers to the following sub questions: What are the significant components of effective learning in early education? How could moral education be enhanced for modern early education curricula? How could

existing technological solutions be improved to incorporate this enhanced moral education curricula, while maintaining effective learning practices for early education students? And, how could this proposed enhanced technological solution be implemented in a responsible manner?

To explore these ideas, I will start with a conceptual investigation in Chapter 2, focusing on early education and moral education. Section 2.1 will define early education and its key cognitive development factors (2.1.1), and explain why it's a crucial area for this research (2.1.2). Section 2.2 will define moral education and explore how it is addressed today (2.2.1), identifying the literature gap lacking focus on "technomoral resilience" and asserting its importance for modern moral education- particularly in relation to early education (2.2.2). Chapter 3 will cover the technical investigation, discussing technology's role in moral education. Section 3.1 will review a case study which uses social robots for general Artificial Intelligence (AI) education for children (PopBots), and Section 3.2 will outline current literature suggestions for using AI for general moral enhancement.

Though a full scale empirical investigation is out of scope for this research project, Chapter 4 will serve as the empirical investigation stage. Section 4.1 will introduce the Value Sensitive Design (VSD) approach, outlining its foundations, methodology, and its pros and cons. Section 4.2 will look at the application of VSD to AI and social robots, addressing challenges like the black box problem (4.2.1) and opportunities for AI for Social Good (AI4SG) (4.2.2). Section 4.3 will synthesize the available literature on full scale VSD investigations surrounding the use of social robots in early education, highlighting stakeholder values and offering solutions where possible.

Due to the limited faculties available in this level of research project, and its researcher, a full scale VSD investigation into the stakeholder values regarding the use of social robots in early education specifically for moral education is not currently feasible. Similarly, neither is an *actual* implementation of this technology for this purpose. It is, however, the hope that this research will prove to be a launching point for this type of

responsible application, providing justification for its creation as well as growth in this field of interest.

Chapter 5 will introduce a proposed project which I refer to as 'MiruBots'. Section 5.1 will provide implementation guidelines which incorporate the cognitive and moral development principles (from Chapter 2), and build on both the PopBots system and the AI for moral enhancement suggestions (from Chapter 3). To paint a better picture of the MiruBots system, Section 5.2 will outline a simple use case for the proposed tool. While the MiruBots system as a technical artifact will be formulated based on the technical investigation of Chapter 3, Section 5.3 will evaluate the proposed solution against the values identified in the conceptual investigation of Chapter 2, followed by Section 5.4 evaluating against the values identified in the synthesized empirical investigation comprising Chapter 4. Altogether, the goal of the proposed MiruBots system is to offer a practical and effective solution for enhancing moral education using technology.

Chapter 6 will wrap up the investigation by recapitulating the results of this research-based exploration, outlining the key takeaways and highlighting the need for responsible implementations of this nature and, subsequently, further research on these innovations. This research is important because it bridges several areas of study which exist separately, including moral enhancement through AI, AI ethics in early education, and the value concerns around using social robots with young students. Currently, no research brings these ideas together into one cohesive direction, nor does any research offer a clear, responsible implementation plan. The goal is to teach future generations about morality and ethical thinking, with a special focus on the responsible creation, use, and impact of technology. And I propose that this can be accomplished through the careful and responsible creation and use of social robots equipped with moral curricula designed specifically for early education.

Using the VSD approach to explore these research goals is particularly relevant because it centers ethical considerations into the design process from the beginning, focusing on the values of all involved, including society at large. Because VSD includes

emphasis on the societal impacts of new technologies, using this framework for specifically this research ensures that the design and implementation of the proposed solution align with core human values. In order to raise future generations to have increased capacity for more effective and peaceful collaboration, values such as empathy, fairness, and moral development are crucial to highlight and exercise in early educational practices, making the VSD approach uniquely advantageous.

This research could have been investigated through the lens of other frameworks or theories, like Kohlberg's Theory of Moral Development, which is a popular reference point within this field (Kohlberg & Hersh, 1977). This theory was criticized by his student Carol Gilligan for lacking a "care perspective" and resulted in her Care Ethics, which instead focuses on the context of relationships in moral reasoning, and has also become a popular reference point in this field of research (Sander-Staudt, n.d., para. 3). Though dealing with moral development and emphasizing empathy, Care Ethics as a focal lens for study is more situated in the interpersonal relationships in a caregiver dynamic, making it more applicable to medical situations than educational ones.

Using these types of theoretical lenses also limits the role of the technology and its development, whereas this research is focused more on practical design requirements of the technological system itself. This is more appropriately addressed through the VSD framework, which provides a structured way to examine the possible solutions of a practical application of this nature. And, by intentionally including stakeholder values early and continuously in the design process, VSD by nature assures that moral and ethical concerns are embedded into the technological solution itself. And, here, this translates directly into the advancement of responsible and meaningful educational outcomes.

Ultimately, I argue for the furthered promotion of critical reflection, open-mindedness, and forethought in society. This can be done by improving both our education system and the way we develop new technologies. We can achieve this by using technologies that are designed with careful attention to human values, using value-focused design techniques. The endeavor to advance our global society towards a

more peaceful and beneficial existence for all its inhabitants is likely to be a never ending effort, and this research aims to serve as yet another stepping stone along that path.

2 Early Education & Moral Education

In order to explore how technology can more effectively enhance moral education in early education, first we must clarify what some of these key phrases mean, namely 'early education' and 'moral education'. This section will provide an overview of these terms within this context, as well as why they are important and what they really entail in the modern day. This section will also identify a gap in the literature surrounding moral education in early education, and suggests filling this gap through the introduction of technomoral resilience as a goal of moral education, underlining the significance of its inclusion specifically within early education and moral education curricula therein.

2.1 What is early education?

Formal education itself varies across the world in terms of starting age, content, structure, duration, and more (Eskelson, 2020). Educational researcher Eskelson discusses further how formal education emerged and is implemented in a variety of ways throughout the globe, but "for universally the same reasons" (2020, p. 29). The goal in its advent, ultimately, was to maintain "state-level societies" by producing citizens who could contribute effectively to society (Eskelson, 2020, p. 33). What this precisely means may differ between societies, but through the synthesis of international educational practices and goals, Eckelson surmises that commonalities found mean that at the end of the day "children in all societies have the ability to learn [...] skills and cultural knowledge through observation, imitation, socialization, and play" (2020, p. 29). This supposition is supported widely in the field of psychology and children's cognitive development, with many

researchers furthering this ideology and basing new studies off it (see: Bandura et al., 1961; Ginsburg, 2007; Tomasello, 1999; Meltzoff, 2007; Vygotsky, 1978; Williamson et al, 2010).

Because of the universality of these key elements of childhood education put forth by Eckelson (observation, imitation, socialization, and play), they will be the primary concerns when discussing early education within this research. And for the purposes of this research, '*early* education' will refer to the earliest levels of formal education through pre-adolescence. Simply for the sake of narrowing the scope within this research, in terms of the US education system this would refer to elementary school and middle school, for example, which encompasses ages within the range of 2-13 ("*A guide to the US education levels,*" 2024).

2.1.1 Cognitive development in early education

Observation, imitation, socialization, and play can be considered the general building blocks with which children experience cognitive development in early education, and this is supported particularly by the work of philosopher and psychologist Vygotsky, who famously studied childhood development and the development of higher psychological functions (Gajdamaschko, 2011). Vygotsky's ideologies stand out when considering important factors for cognitive development in early education because of its socially grounded approach to childhood learning.

Though there are other theories and psychologists which could provide a basis for establishing key factors in early cognitive development, they tend to overlook the socio-cultural factors which are highlighted in Vygotsky's conceptualizations and rather focus on individual and internal cognitive changes. While generally important, these factors are less directly related to the interactive nature of exploring the potentials of social robots in early education. Thus, Vygotsky's framework aligns more readily when considering the factors necessary for this research.

Additionally, Vygotsky's ideas tend to blend some more popular psychologist's cognitive development understandings like Piaget or Erikson (see: Piaget & Inhelder, 1969;

Orenstein, 2022), just in a more positive, constructivist light which lends itself more applicably to integrating technology into early education practices. Some key cognitive development concepts from Vygotsky which will be important considerations in this research include (i) culture-specific tools and (i.a) scaffolding, (ii) the zone of proximal development, (iii) the dialectical method, and (iv) private speech.

Vygotsky's research in cognitive development suggests that cognitive development is strongly impacted by social and cultural factors, emphasizing social interaction's role in developing mental abilities like reasoning and 'making meaning' (Mcleod, 2024). According to Mcleod (2024), Vygotsky believed that "cognitive development is a socially mediated process in which children acquire cultural values, beliefs, and problem-solving strategies through collaborative dialogues with [...] more knowledgeable other[s]" (p. 2), which typically referred to parents or teachers. Though, in Chapter 5 we will explore the possibility of this 'more knowledgeable other' (MKO) as an AI-powered social robot. The key takeaway from this understanding of cognitive development within the context of this research is that, essentially, children learn values and morals through experiences and interactions with MKOs (such as by observing, imitating, socializing, and playing with them).

In conjunction with this emphasis of social influence on cognitive development, Vygotsky's psychology cited (i) culture-specific tools as "reflecting [the] socially constructed ways in which society organizes the various [...] tasks faced by a growing child and the physical and mental tools that society provides [...] to master those tasks" (Gajdamaschko, 2011, p. 696). In the case of this research, these tools will include the suggestion of social robots as physical tools, which will themselves provide new forms of scaffolding as mental tools. (i.a) Scaffolding for Vygotsky in this context refers to the 'support structures' provided by the MKO which help the student to understand the task and strategize how to find a solution, gradually allowing them to gain independence on the task as their confidence grows with their gradually improving problem solving skills (Schaffer, 1996). A simple example of scaffolding could be explaining to a child the desired

result of an unfinished puzzle and then showing them how to test out different sides of a piece until it fits with its neighbor, so the MKO here is providing parameters and an example of a problem solving strategy (Schaffer, 1996).

Scaffolding is one tool MKO's themselves can use to help aid in children's cognitive development of their own forms of scaffolding, and this is particularly handy when MKO's consider the student's zone of proximal development (ZPD). (ii) The ZPD refers to the gap between what a student can learn by themself, what is above their ability to comprehend, and what they can learn with the help of an MKO or with technology (Wass & Golding, 2014). In the case of this research that would mean the same thing in one, a technological tool which is more knowledgeable than the student *and* actively helps them learn through the use of scaffolding techniques which are developmentally appropriate to the student.

Another facet of Vygotsky's cognitive development theories is (iii) the dialectical method, which stems from Hegel's dialectics, or the process of contradictory views attempting to converge to a truth through argumentation and reasoning (Maybee, 2020). This in itself is reminiscent of the Socratic method, which will be germane also in Chapters 3 and 5. For Vygotsky, his dialectical method incorporates the constant changing nature from Hegelian dialectics, that the interactions and interrelations between the contradictory aspects of things are the driving forces of development (Gajdamaschko, 2011). In essence, "development [is] viewed as constant transformation" (Gajdamaschko, 2011, p. 697). This will be a relevant understanding when discussing moral education and technomoral resilience in later sections.

The final pertinent concept of cognitive development in children from Vygotsky is 'private speech'. Through the use of scaffolding from an MKO when addressing a student's ZPD, the MKO is helping develop the student's private speech (Vygotsky, 1987). In other words, by doing things like talking through a problem or showing a child a new technique for solving a problem, an educator is helping a child formulate the 'how to do it' part of the task for themself, which they can then use later when encountering a similar problem. This results in a dialog that the child uses *to* themself and *for* themself, and not to or for

another person, thus, 'private' speech. This occurs typically from the age of three, and transforms into silent internal speech for self-regulation by the age of seven (Mcleod, 2011, p. 8). Encouraging and developing private speech will be discussed further in Chapter 5 as an important contribution of social robotics for enhanced moral education in early education.

In sum, for the purposes of this research, 'early education' refers to the formal education a child goes through until they reach adolescence which encompasses the cognitive development of the child's skills and cultural knowledge through their experiences of observation, imitation, socialization, and play. Through these types of experiences, a child explores and expands their zones of proximal development from a more knowledgeable other's guidance using culture-specific tools, commonly including the dialectic method and other variations of scaffolding techniques.

2.1.2 Why emphasize early education?

Within this research and with regard to enhancing moral education, the ultimate goal here is to make as big an impact individually and societally as possible in terms of producing responsible citizens who reflect on their actions and try to make decisions which mitigate harm done to themselves and others. While this goal can and perhaps should be addressed and evaluated at all points during one's lifetime, beginning to instill a sense of morality in early education may prove to be particularly effective in ensuring moral values are truly instilled in people as they grow up to become impactful members of society.

Studies show that high quality early education programs have a significantly positive effect on later academic performance (Barnett, 1995), emotional skills (Weiland & Yoshikawa, 2013), and even socio-economic status (Ritchie & Bates, 2013)- which all affect one's ability to contribute to society, ultimately. There are of course studies showing the difference between those from impoverished areas without early intervention and those with, proving that typical education alone without early developmental intervention is less effective for cognitive and social development (Campbell & Ramey, 1994). Additionally, it is

actually more economically efficient to invest in educational equity and high quality early education programs which promote the development of cognition and character (Heckman, 2011).

There is evidence to suggest that adversity in early childhood can "weaken developing brain architecture," which affects not only long term academic performance but also long term health, which again inherently affect future societal contributions (Harvard University, 2007, p. 1). There is, however, also evidence to suggest that early intervention can prevent these negative consequences, and the earlier the better (Harvard University, 2007, p. 2; Guerra & Bradshaw, 2008). So through emphasizing the enhancement of moral education specifically in *early* education, we can actually help children receive the tools they need to deal with various kinds of adversity they will encounter all throughout their lifelong development. There is further research which highlights that building a child's resilience, or the “capacity to remain flexible in our thoughts, feelings, and behaviors when faced by life disruption” (Pemberton, 2015, p. 2), through scaffolding learning techniques as early as possible is also a key factor to later success in both mental capabilities and physical health (Harvard University, 2015). This emphasis on the importance of resilience will be reinforced when discussing technomoral resilience later on.

In brief, early education is an important area to target in studying the enhancement of moral education because of its powerful impact within human cognitive development. This imminently shapes the lives and realities of individuals and society as a whole. Though emphasizing early education is not the singular solution to creating a better world with more responsible citizens, instilling these values early increases our chances of this moving forward. If we can't improve everything in a child's environment as they grow, we can still take a positive step. By giving them tools early on, we help them cope and move forward more effectively, even in challenging or less-than-ideal environments.

2.2 What is moral education?

As we have ascertained the crucial elements of early education and the significance of providing a high quality one (which includes moral education and resiliency), we can now decipher what 'moral education' more characteristically entails in present day early education. Traditionally, moral education has been a fundamental concern within the philosophy of education, essentially asking the question 'how do we learn how to be good and moral humans?' and is interchangeably referred to as 'values-,' 'ethics-,' or 'character-' education (Chazan, 2022, p. 23). In the past this was of course linked largely to religion and religious studies, but that has since changed in the modern age as moral and ethical education are now simply inserted into general education practices unrelated to religious studies (Chazan, 2022, p. 24).

For the purposes of this research, the primary interest of 'moral education' as a concept is the way in which children are learning how to deal with 'moral situations.' These are situations in which one must make a decision on what to do, typically between conflicting values (Chazan & Soltis, 1973). So, instead of choosing between right and wrong, moral conflict is rather about choosing between two rights or two wrongs (Chazan, 2022). This is difficult for even adults to accomplish in many cases, so for children in early education this must of course be scaled down to age appropriate moral dilemmas for consideration.

Furthermore, moral education is effectively fostered through developing a child's reasoning skills (Meyer, 2023) most commonly done so through peer discussion (Kohlberg, 1984). This method of cultivating moral development is both complementary to and supported by Vygotsky's dialectic method from Section 2.1 and the understanding that cognitive development itself is driven by the constant transformations that occur via discussions which entail reasoning through contradictory notions (Gajdamaschko, 2011).

Additionally from the discussion on 'early education' in the previous section, we know that children are learning their values, beliefs, and decision-making strategies through their collaborative dialogues and experiences with the people and things they

observe, imitate, socialize, and play with in their lives (Mcleod, 2024). So, even when not explicitly being thoughtfully instructed on values and ethical thinking, children are still absorbing a great deal which will go on to inform their beliefs and eventually their decisions, good or bad. This influence can come from anywhere and anything, like the technology they are interacting with and all that that may bring with it, which is why modern moral education has evolved to include topics like AI ethics (Yang, 2022; Adams et al., 2023).

2.2.1 AI ethics in early education

AI is a complex technology which most adults struggle to understand, and its implementation and use present issues, especially for children. Issues like biased AI results can shape children's understandings of the world in incorrect or unfavorable ways, which is why AI ethics is important to include in early educational practices. By educating children on how AI works and why they should be critical of it because of its issues, children can grow up more aware and use the technology more responsibly with fewer instances of undue influence.

So, not only has general education added digital literacy to the roster, with support and advocacy from the UN (UN CRC, 2021), for example, but the curricula is also moving specifically towards AI literacy, as it has been established that kids can and do effectively interact with and absorb age-appropriate knowledge of this type (Touretzky et al., 2019). And since children are growing up with this technology, with all of its pros and cons and all of the moral dilemmas it presents, this AI literacy education is now necessarily incorporating AI ethics, at all levels of education (Adams et al., 2023; Aitken & Briggs, 2022; Williams et al., 2022).

This AI ethics curricula includes the introduction and examination of notable values and concepts such as privacy, accountability, safety and security, transparency and explainability, fairness and non-discrimination, human control of technology, and promotion of human values (Aitken & Briggs, 2022, p. 5). Touretzky et al. (2022) provide a

more practical example of what this actually looks like in early education, “students in grades 3-5 should exhibit critical thinking about the impacts of new AI applications, e.g., self-driving cars will be a boon to people who cannot drive themselves, but may also put taxi drivers out of work” (p. 9798).

So not only are children in early education learning ‘how to be a good person,’ but they are also learning ‘how to be a good person’ in relation to technology, like how to create and use technologies like AI responsibly and critically— all as part of their moral education within their general education. Something that is missing from much of the literature discussing this, however, and one gap that I attempt to fill, is introducing the concept of ‘technomoral resilience’ into moral education in early education.

2.2.2 Technomoral resilience in early education

As previously established, resilience is itself already an important life skill for cognitive and emotional development (Harvard University, 2015). Resilience science has a combinatory background rooted in developmental psychology and supplemented by biological and medical studies, giving the importance of this life skill both normative and empirical backings (Masten & Barnes, 2018). Resilience as a concept has also been applied to morality, resulting in ‘moral resilience,’ or the ability to respond flexibly to the changing nature of moral norms and values within society (Swierstra, 2013).

Bauer and Hermann (2022) have taken this one step further and urge the application of this concept towards society’s ever changing moral norms and values, particularly as they are influenced (at least in part) specifically by technological developments— hence ‘technomoral resilience.’ A predominant concern of technomoral resilience is the movement from stabilization, to destabilization (of one’s morality due to technological influence), and back to a state of stabilization (Bauer & Hermann, 2022, p. 64). They also highlight that this building of resilience to moral norm fluctuation caused by technological advancements does not exclude resistance to these changes, but rather

incorporates a healthy balance of both critical evaluation and adaptability (Bauer & Hermann, 2022, p. 65).

They propose that a goal of moral education ought to be bringing technomoral resilience into the equation through fostering (i) “moral imagination”, (ii) “a capacity for critical reflection”, and (iii) “a capacity for maintaining one’s moral agency in the face of disturbances” (Bauer & Hermann, 2022, p. 59). In order to develop (i) moral imagination, it is suggested to develop vocabulary which can help describe experiences and emotional reactions, and to play through imagined scenarios or even use video games to simulate and more fully experience these imagined scenarios of future technologies’ possible influence on morality (Bauer & Hermann, 2022, p. 68).

The process of (ii) developing critical reflection is proposed through critically evaluating scenarios from both imagined and real technology’s moral impact(s) from a variety of ethical perspectives, questioning also the “fundamental moral concepts that are involved [...] and [the] potential changes of these concepts” (Bauer & Hermann, 2022, p. 68). And lastly, (iii) maintaining moral agency is to be developed through the repeated rehearsal of the previous suggestions, with an emphasis on learning to manage emotions and foster self-questioning as well as open-mindedness (Bauer & Hermann, 2022, p. 69). Chapter 5 will explore the use of social robots to accomplish all three of these steps within early education.

The capacity building approach suggested by Bauer and Hermann (2022) is echoed by Christen and Narvaez (2012), who also put forth that "having a set of ethical capacities honed to automaticity [...] makes it more likely that an individual will act virtuously" (p. 25), which is the goal at the end of the day, for enhancing moral education and, hopefully, achieving a more harmonious world. They reinforce the supposition that early childhood is the optimal starting point for moral education intervention, and that it can be “culturally sensitive and tailored to the needs of the individual” (Christen & Narvaez, 2012, p. 26). This will also be explored further in Chapters 3 and 5 through the use of AI-powered social robots. Moreover, ethical and moral education has been and should continue to be a

pluralistic endeavor, as in, children should learn from as many responsible sources as they can, with as much collaboration as possible from these sources, for the development of effective moral education (Burroughs, 2018).

To recap this section and chapter, moral education is concerned with the ways in which we learn how to be good people. In the progressive era in which we live, this has come to inherently include technology and its impact on both society and the moral norms of the day. It is important for new generations to remain critical yet adaptable in their moral values. As technological advancements bring new realities and values, this ability can be transformative.

Children need to learn how to responsibly create and interact with technologies like AI as they develop mentally, emotionally, and physically. Equally important is that they learn to co-evolve responsibly with these technologies. This includes understanding the new challenges and moral dilemmas that come with these advancements. Ultimately, this will help shape society in unprecedented ways. The best chance we can give the future generations involves responsibly and thoughtfully starting this technomoral resilience enhanced moral education as soon and as effectively as possible within early education. So, in the interest of exploring the responsible use of technology for enhancing moral education in early education, we now understand better what the significant components of effective learning in early education are, as well as how moral education can be enhanced for modern early education curricula.

3 Early Education & Moral Education— Using AI & Social Robots

With a more comprehensive understanding of early education, moral education, and the importance of both and what they ought to include in the modern day, we can take the next step in examining *technology's* role in the endeavor of effectively enhancing moral

education in early education. This chapter will explore a few uses of technology which will both ultimately contribute to an actionable solution proposal for consideration while moving forward in this effort. These technological uses include an actual implementation which demonstrates technology as an aid in early educational endeavors, followed by a theoretical implementation which demonstrates technology as an aid in moral enhancement endeavors. In combination, these implementations provide a possible way forward for technology to be used effectively for moral education in early education.

More specifically, in this chapter I will first outline a case study, the MIT PopBots, which has already successfully employed social robotics in early education. The MIT PopBots project was created for the purpose of teaching young kids about general AI concepts, as well as for introducing them to AI ethics concepts. Then, I will outline a prospective technological implementation found in the literature from research concerning the use of generic text-based AI applications for moral enhancement. These two separate use cases, social robots in early education and AI for moral enhancement, provide the technical investigation of the VSD approach in this project and will lay the foundational groundwork for the enhanced solution I will propose more fully in Chapter 5.

3.1 The MIT PopBots case study

In 2019, research team Williams et al. published work outlining their PopBots system. They acknowledged that young children are now growing up interacting with AI technologies more and more commonly, without fully understanding AI concepts nor the dangers they may pose. This spurred them to develop a Preschool Oriented Programming (PopBots) Platform, created to teach young children about AI concepts in an empowering way which shifts them from passive users to active creators who understand AI technology foundations (Williams et al., 2019a, p. 9731). The project does this by leveraging a social robot as a learning companion, allowing children to build, program, train, and interact with the robot in order to grasp fundamental AI concepts.

Figure 1 from Williams et al. (2019, p. 1) visually shows the PopBots system in its entirety, comprising a smart phone with an emotive bot face situated atop a LEGO body (designed uniquely by the child and/or parent), along with a tablet which allows the child to dynamically interact with the PopBots system's curriculum activities. The child interacts with the bot mainly via the tablet, completing activities on the tablet's interface and receiving verbal and visual feedback from the bot. Through a carefully thought-out age-appropriate curriculum, the children are able to gain experience learning about complex AI concepts *while* interacting with an AI system in real time.

Williams et al. (2019, 2019a) found that their PopBots system fostered computational thinking and digital literacy in an engaging way, and they were able to curate a curriculum which effectively accomplishes this through principles like hands-on learning, transparency and tinkerability, and creative exploration. The curriculum topics successfully taught young kids about knowledge based systems, supervised machine learning, and generative music AI (Williams et al., 2019a, p. 9731). Through empirically evaluative assessments, they found that children as young as four years old were successfully grasping these AI concepts.

The use of a social robot as a tool to aid in early education practices proved to be particularly effective, as it helped demystify AI and made abstract concepts tangible and relatable for the children (Williams et al., 2019, p. 9; Williams et al., 2019a, p. 9730). Furthermore, the project highlighted the potential associated with using hands-on, interactive tools to teach complex subjects to young audiences. The success of PopBots suggests that children are capable of understanding and engaging with AI technologies when these are presented in an accessible and age-appropriate manner (Williams et al., 2019, p. 10).

An additionally interesting finding from this case study was that they found that the activity they created which essentially asked the child to step into the mind of the AI robot (e.g., if this is the scenario with inputs x, y, and z, how would the bot behave?), actually boosted the children's Theory of Mind skills (Williams et al., 2019, p. 10). Based on typical

Theory of Mind skills stages and assessments, the younger children in these studies should not have been able to achieve the level of perspective-taking they did after participating in the study (Barnes-Holmes et al., 2004).

The particular Theory of Mind skills that children need to grasp in order to understand AI include understanding knowledge access (the awareness that someone else might not know something you know), understanding content false belief (the recognition that another person may hold an incorrect belief that influences their actions), and understanding explicit false belief (knowing how a character will act based on its perceived knowledge) (Williams et al., 2019, p. 2). Through assessments before and after the children participated in the PopBots activities, Williams et al. (2019) concluded that not only were the PopBots activities successful in educating the children about AI concepts, but by going through these hands-on activities with social robots in educational settings the children actually gained more ability to open their minds up and take on other perspectives.

Later work from Williams et al. (2022) found that a combined AI and ethics curriculum allowed middle schoolers to effectively grasp technical concepts while developing a critical lens as to how the technology impacts society, though this was not accomplished via the use of social robots specifically. Still, this encourages a two birds-one stone mentality that kids may more effectively absorb these technically and emotionally complex concepts when discussed in tandem. This is a beneficial finding to keep in mind when designing systems and curricula with the goal of effective moral education.

Overall, the successful elements of the PopBots case study in terms of this goal align greatly with the significant components of effective learning in early education, as identified in Chapter 2. Effective learning practices for early education students incorporate observation, imitation, socialization, and play— and these are the exact types of experiences that can be accomplished with social robots. A technology with this effective functionality in an educational setting could be enhanced with an age appropriate

moral education curriculum, and this would be one possible avenue in which AI could be designed to aid in this endeavor.

3.2 AI for moral enhancement

While the PopBots case study is informative in terms of exploring an effective technological aid for early education students, it does not by itself tell us about the use of technology for *moral* education. To investigate the use of text-based AI technologies (as opposed to social robots), this sub section will explore the best practices accumulated by Volkman and Gabriels (2023) in using AI for moral enhancement. Moral enhancement typically refers to moral development which occurs in a neuroscientifically informed manner, often through pharmaceuticals (Christen & Narvaez, 2012). Obviously, using pharmaceuticals to enhance the moral development of children is not an ideal route, so a capacity building approach is again preferable (Christen & Narvaez, 2012), and Volkman and Gabriels (2023) outline a path forward which instead utilizes AI technologies to do so.

Their proposal builds upon Lara and Deckers' (2020) idea of AI as a "Socratic Assistant," where AI is envisioned as a tool that can engage users in reflective and critical thinking about moral issues, helping them to better understand and refine their moral beliefs and behaviors. Taking this role of AI as a socratic dialogue partner one step further, Volkman and Gabriels (2023) suggest expanding this idea in such a way that a socio-technical system is formed which collectively facilitates moral engagement.

This socio-technical system is an important distinction from simply the 'Socratic Assistant' since, as Volkman and Gabriels (2023) identify, the model of moral engagement which humanity has adopted thus far, and which has proven to be both important and successful, comprises a socio-technical system of its own accord. Through books, symposiums, online forums, etc., humans have discussed and debated moral issues, garnering clearer understandings through these kinds of socio-technical systems (Volkman & Gabriels, 2023).

They propose that these systems are to be composed of several AI interlocutors, each trained in a specific ethical tradition and therefore able to engage the user in a variety of viewpoints in an active discussion. Volkman and Gabriels (2023) liken this environment to an interactive book with multiple characters engaging with the reader based in viewpoints like Stoicism and Buddhism, amongst other ethical perspectives, where the characters' collective goal is to help the reader cultivate their own wisdom.

They criticize previously suggested AI implementations for moral enhancement in that they focus on AI providing the 'right' answers, thus, "ultimately reducing morality to the output of some algorithm" (Volkman & Gabriels, 2023, p. 3). Not only is this problematic for allowing technology to determine the 'correct' moral position to take in something, thereby relinquishing human agency and granting perhaps too much power to technology, but this is also problematic for actively sabotaging the development of human capacity for critical reflection (Volkman & Gabriels, 2023).

They also cite past suggestions in this regard as constituting machines as being "morally superior to humans," emphasizing again that these tools should instead provide "auxiliary enhancement" and not dictatorial engagement (Volkman & Gabriels, 2023, p. 7). Additionally, they caution against the other suggestions for their goal of providing 'ease' in moral endeavors, as technology is often developed to do. They argue instead that this more complex system which embraces Socratic methodologies is more akin to the appropriate role of a philosopher- to pose the hard questions and to make one think and reflect and grow critical capacities for practical wisdom (Volkman & Gabriels, 2023, p. 12).

In toto, the benefits of using AI for moral enhancement, as suggested by Volkman and Gabriels (2023), include several key points. First, AI offers a non-invasive approach, unlike other neuroscientific interventions which are popular in modern moral enhancement practices. Second, interactive 'AI mentors' could be more effective than books or other media aimed at moral development. Lastly, instead of simply giving the 'right' answers, 'AI mentors' could help build users' capacity for critical thinking through gentle guidance and reflective discussions.

This is the most important takeaway when considering the research goal of exploring how AI can help in providing an enhanced moral education curriculum, as this translates directly into an effective technological design requirement. Designing the use of AI in this way also aligns with the capacity building approach recommended for both cognitive development in early education and fostering technomoral resilience in moral education (as discussed in Chapter 2).

With a more solid understanding of a current successful solution in early education (PopBots) and a well thought-out theoretical solution for moral enhancement ('AI mentors'), it is more conceivable how these technologies may be combined into one solution which could effectively aid in enhancing moral education within early education practices. The technological solution proposal pulling from these two ideas will be further fleshed out in Chapter 5.

In brief, this Chapter's technical investigation into the use of AI and social robots has supported that, when thoughtfully improved and designed to incorporate an enhanced moral education curriculum, AI technologies could provide a useful, interactive tool which maintains effective learning practices for early education students. The next step is to investigate further how this technological intervention could be implemented in a responsible manner.

4 Value Sensitive Design

In the perpetually evolving landscape which encompasses science, technology, engineering, and the increasing interrelations of these fields, the integration of human values into the design and development processes of the complex innovations they produce has become increasingly crucial. Value Sensitive Design (VSD) is an approach which aims to ensure human values are at the forefront of each step of the design and implementation processes for these labyrinthine inventions, with all of the varying

concerns they may present. Exploring the use of technology to enhance moral education in early education *responsibly* inherently begets the careful consideration of ethical and moral values, so the VSD approach is exceptionally pertinent to consider

Originating in the 1990s, VSD was a response to the need for integrating human values into technological design, rather than treating them as secondary considerations (Friedman, 1996). This framework was coined by Batya Friedman as part of her research into the ethical implications of technology, especially in terms of how design could actively influence social values (Friedman, 1999). The approach has evolved to provide a robust methodology for incorporating ethical and social considerations into the design of technological solutions like information systems and other computer and digital technologies (Friedman, 1996).

Though based in the computer sciences, it can also be found in explorations related to biotechnology and health sciences, sociology, philosophy, environmental sciences, and transportation sciences (Winkler & Spiekermann, 2018). And, although this approach provides a framework for incorporating ethical values into the design process, it does not provide an algorithm for making the 'right' decisions within technological design and it does not guarantee a wholly ethical product as a result of its use. These are of course outcomes realistically reliant on the diligence and thoroughness of the designers employing the approach.

VSD's approach to address the complex interplay between technology and human values is a systematic and principled one, and this chapter delves into the theoretical foundations, methodological approaches, and practical applications of VSD. This will be accomplished by drawing insights from key sources. First, by synthesizing a more in depth explanation of the approach as applied to generic technological developments, including further discussion of its critiques and limitations as a framework. Then, through the examination of the framework applied specifically onto AI and social robots as the technology of concern. And finally, by looking at the applications of VSD in practice

through an analysis on actual studies which explored the use of specifically social robots expressly in early education.

4.1 The Value Sensitive Design approach

The following section will provide a closer look at Value Sensitive Design as a framework for approaching general technological development with values, including values with moral import, as the driver of the design and implementation processes. First I will explain further the theoretical foundations of VSD, as well as how they complement the ideologies emphasized throughout this research. This will be followed by the more in depth elucidation of the methodological approaches VSD employs, and finally this section will close with an overview of both the pros and cons associated with the use of this framework.

4.1.1 Theoretical foundations of Value Sensitive Design

As defined by Friedman, Kahn, and Borning (2002), VSD is rooted in, amongst other things, the belief in proactive impact, that it is prudent for effective impact to be made through being proactive in influencing the design of technology starting early in, and also throughout, the process of design and implementation of technologies (p. 2). This is complementary to my ideology discussed in Section 2.1.2 emphasizing moral education in early education, as opposed to the continued efforts of this endeavor in higher education- that greater societal impact can more beneficially be made at this stage of development.

Of additional importance to VSD theory is the understanding that societal values are not determined by technological developments, nor are technological developments deterministic of societal values; rather, VSD embraces these relationships as interactional, with each shaping the other (Friedman et al., 2002, p. 2). This in turn allows for the constant valuation and incorporation of the changing landscape of values and their relation to technology, resulting in the highlighted feature of an iterative design. Because of the volatile nature of societal values and how technological developments may impact them, and vice versa, VSD encourages an iterative process within design and

implementation, continuously accounting for whatever fluctuations may arise as uses and usership progress (Friedman et al., 2002, p. 2).

This element of VSD's theoretical philosophy is reminiscent of that of technomoral resilience, as discussed in Section 2.2.2. It provides the understanding that individual and societal values are in constant transformation and are influenced by technological advancements. Therefore, in moral education we must account for this variance and build capacity for resilience in the face of these types of disruptions and alterations. So, too, must researchers who take on this design approach be thoughtful of these fluctuations and ensure their incorporation into the design considerations of the technology at hand, ensuring the iterative nature of VSD is upheld for this purpose.

Furthermore, VSD is committed to the incorporation of values such as privacy, autonomy, and trust into technology design. It stresses that these values are not merely byproducts or afterthoughts, but are absolutely essential components of technology that shape both user experience and the impact that the technology will have on society as a whole (Friedman et al., 2002, p. 3). Because VSD gives power to values stemming from moral epistemology, it ensures certain values are upheld universally, regardless of any subset of stakeholder desires (Friedman et al., 2002, p. 2). VSD has more recently evolved to explicitly include the irrefutable maintenance of at least three universal values: human well-being, justice, and dignity (Friedman & Hendry, 2019).

The theoretical foundations of VSD were elaborated by Friedman, Kahn, Borning, and Huldtgren in 2013, who acknowledge the importance of considering both direct and indirect stakeholders in the design process. Direct stakeholders are those who interact with the technology, while indirect stakeholders are those affected by the technology's use. This dual focus ensures that the broader societal implications are considered alongside individual user needs (Friedman et al., 2013, p. 66). And, of final importance here, VSD operates on three main types of investigations: conceptual, empirical, and technical (Friedman et al., 2013, p. 59)— and these will be explained in the following sub section.

4.1.2 Methodological approaches in Value Sensitive Design

The methodological framework of VSD is multifaceted, involving a three part approach consisting of conceptual, empirical, and technical investigation techniques (Friedman et al., 2013, p. 59). These investigations are iterative and integrated throughout the design process to ensure that values are continuously addressed and refined. This approach contrasts with traditional design methodologies, especially within computer and informational sciences. These design approaches often treat ethical considerations as an afterthought and may not even consider multiple iterations for the purpose of more effectively addressing ethical concerns (Zunger, 2018).

4.1.2.1 Conceptual investigations

Conceptual investigations are foundational to VSD, involving the identification and articulation of values that are relevant to the technology and the stakeholders involved in the creation and use of the technology, which include both the direct and indirect stakeholders (Friedman et al., 2002, pp. 2-3). Friedman et al. (2002) continue on that the aim of the conceptual investigation phase is really to work towards understanding the ethical implications related to the technology and its use through mapping out and formulating the real and/or projected values that may arise from its use.

This means explicitly defining stakeholder values as well as analyzing any conflicts or tensions (or even possible alignments) that could emerge in weighing these values against each other (Friedman et al., 2013, p. 60). An example of this could be when investigating the use of technology for moral enhancement— some conflicting values might include weighing privacy as a value from a stakeholder such as a regular user (who wants to know that their personal data is not misused), against transparency as a value from a stakeholder such as a tech savvy user (who wants to know how the technology works, including data collection and decision-making processes).

4.1.2.2 Empirical investigations

Empirical investigations in VSD involve gathering data from stakeholders to understand both their values as well as how they interact with the technology (Friedman et al., 2002, p. 3). Friedman et al. (2002) expand that this investigation method involves both qualitative and quantitative research methods, which include techniques like surveys, interviews, observations, and ethnographic studies in order to gather insights from stakeholders and to help designers understand how values emerge in real use cases in the world.

At their core, empirical methods employed in the VSD approach bring to light the human context in which the technology functions, and they also uncover the (sometimes unpredictable, and) nuanced ways in which users experience, perceive, and prioritize different values as a result of the technological use (Friedman et al., 2013, p. 61). This ultimately allows designers to discern users' expectations, needs, practices, and values, in order to make more informed decisions throughout the technology's design process (Friedman et al., 2002, p. 3).

An example of an empirical investigation could be a research team conducting interviews and distributing surveys to key stakeholder groups like parents and educators, in order to gain more concrete insight into the needs and values that may arise (and which could prove to be in conflict) when considering the use of AI for moral enhancement in early education. Friedman et al. (2002) also stress the need for designers to be aware of the complex relationships that present themselves between specifically usability and human values which focus on ethical importance, that sometimes they may support each other and yet at other times they may need to be scrutinized deeply for the appropriate balance of give and take necessary for the creation of a viable product which still upholds moral values.

4.1.2.3 Technical investigations

Technical investigations in VSD focus on the actual design and implementation of the technology of concern in a manner that supports the values of importance which were

identified in the conceptual and empirical investigational methods (Friedman & Hendry, 2019). This translates to, where applicable, the development of prototypes of the technology, as well as conducting usability testing to analyze whether the technology and its use align with the values identified for consideration, or whether some values were hindered through the design process and did not make it into the product prototype (Friedman et al., 2002, p. 3).

Friedman et al. (2002) continue on to discuss that the nature of technical investigations differ from the other investigational methods because this method focuses on evaluating the technological features themselves and how the technology is used, as opposed to focusing on the stakeholders. This methodological phase does, however, stress and provide a good jumping off point for the iterative element of the VSD approach.

After examining the trade-offs and impacts of design choices in the use of the prototype, or even in the use of related technologies when a prototype is not yet viable, new prototype designs, or a new design of systems, can be drawn up again based on feedback and observations in order to more effectively address or incorporate the values of import (Friedman et al., 2013, p. 61). An example of a technical investigation for examining the use of AI for moral enhancement in early education could be training an AI model in a specific ethical framework, like Kantian or deontological ethics, and providing it with a set of age appropriate moral dilemmas to socratically talk through with a student in early education, and then observe and evaluate the implementation and its use as it pertains to the values of significance.

4.1.3 Weighing the value of Value Sensitive Design

The Value Sensitive Design approach provides a framework for putting ethical considerations in focus while designing technology that shapes society, and this is a much needed tool which in itself is beneficial in the world. However, there are still critiques to consider when engaging with this tool.

The VSD methodology is proven robust and widely useful through its emphasis on iterating the investigational methods outlined in the previous sub section, as well as through underlining that the VSD approach can begin at any one of those investigative phases and subsequently built upon in response to the findings from the previous phase in which the research team engaged (Friedman et al., 2013, p. 59). This dynamism is paralleled with the dynamic nature of values, which are mutable and can vary between contexts as well as cultures.

This was one of the main points of critique from Manders-Huits (2010), who argued that much of the VSD framework was 'nebulous' and lacked concrete parameters. By 2017, VSD had itself evolved and, along with Van Der Hoven, Manders-Huits put forward an updated view of the framework which incorporated the understanding that in order to remain relevant and effective, VSD methodologies must adapt through ongoing assessments addressing this variability.

Furthering this lack of solid guidelines, another critique argued that VSD does not provide any solution to the challenge of translating values into operationalized design requirements. This brings up a good point, this issue is not addressed explicitly within this framework, though, perhaps the answer to that is not necessarily under the purview of the VSD approach. Friedman and Hendry (2019) put forth that the way forward when considering concrete design requirements which effectively apply value considerations requires interdisciplinary collaboration among the designers, engineers, and ethicists.

Another critique of VSD touched upon by Borning and Muller (2012) is that although VSD is itself valuable, its incorporation into mainstream practices will not be accomplished without institutional support and broader educational initiatives. A value-sensitive and ethically focused culture in technology development will not be fostered without appropriate VSD (and similar) training being made available to engineers and designers.

Despite these critiques, the VSD approach provides a principled framework for analyzing and integrating human values into the design and development of technology.

Through its theoretical foundations and its employment of conceptual, empirical, and technical investigations with ethical considerations at their core, VSD ensures that technologies are not only functional, but that these innovations are also socially responsible and ethically sound. Its holistic integration of values, involvement of diverse stakeholders, and iterative methodologies make VSD an advantageous approach compared to traditional design frameworks in creating technologies that are accepted, trusted, and aligned with the values and needs of users and society at large.

4.2 VSD applied to AI & social robots

Creating and implementing technologies like AI and social robots in a socially responsible manner presents its own set of ethical and moral concerns, and this has brought specific attention to the use of the Value Sensitive Design approach in both of these overlapping domains. Technology researcher van de Poel (2020) brings to light that the VSD approach as it is alone does not sufficiently address the ethical concerns for the ethical design of AI and the underpinning machine learning (ML) models, and subsequently, any and all applications which may utilize these technologies would also be compromising their ethical assurances. This criticism stems largely from the issues that arise due to the black box nature of AI and the ML models underlying these types of programs and applications. The black box problem and the explainability problem with AI technologies and their uses is not a new topic, and has resulted in initiatives like explainable AI (XAI) (Ali et al., 2023).

In response to this criticism and gap in the framework, van de Poel teamed up with fellow researcher Umbrello in 2021 to propose a modified version of the VSD framework which aims to address this type of issue, as well as other issues, that the standard VSD approach does not account for when it comes to AI technologies. This updated approach specifically incorporates the Artificial Intelligence for Social Good (AI4SG) initiative's five ethical principles and seven essential factors for ensuring ethical values are accounted for and integrated into design requirements for AI technologies (Floridi et al., 2018). The

updated VSD approach for AI also pulls from the European Commission's (2019) four ethical principles laid out by their High-Level Expert Group on AI.

In order to gain a more complete understanding of the bigger picture of the application of the VSD approach onto the design and implementation of AI and related technologies, this section will first give an overview of the black box problem with AI as well as its current solution in progress, XAI. Following that, we will delve deeper into the AI4SG initiative and its five ethical principles (Floridi et al., 2018) and seven essential factors (Floridi et al., 2020), and, in addition to that, an overview of the EU's four ethical principles (European Commission, 2019). And finally, to wrap up this section we will go through Umbrello and van de Poel's (2021) proposed solution for including AI specific values in the application of the VSD framework onto AI and similar technologies, as well as a swift synopsis of the findings and suggestions from the Schmiedel et al. (2022) research team's discussion of VSD for use specifically when applied to the responsible development of social robots.

4.2.1 The black box problem & explainable AI

A commonly discussed issue within AI is considered 'the black box problem', in which it is unintelligible how the underlying programmed algorithm is functioning and producing the results that it does, resulting in a lack of trust in the system and the technology as a whole (Zednik, 2019; von Eschenbach, 2021). For further explanation, this nomenclature is in relation to black box testing in the field of software engineering, where the programmer has certain use cases to satisfy so they set up a testing suite which outlines the expected output based on a set of known inputs, without going into detail *how* the program accomplishes these outputs (Nidhra & Dondeti, 2012).

Similarly, as von Eschenbach (2021) expounds, when it comes to AI, the inputs may be known and the results based on those inputs may be known, but exactly how and why the algorithm is making its calculations is not yet fully comprehensible- even for experts in the field. This uncertainty of the inner workings of the program's algorithm results in

the system being described as opaque, and as lacking transparency (von Eschenbach, 2021). Scientists and researchers Campolo and Crawford (2020) even referred to this type of 'mystifying' technological computation as enchanted determinism, likening the emergence of AI to the advent of alchemy, as people could not explain their early forms of chemistry and why their 'magic' 'potions' worked.

To counter these issues of opacity and transparency from enchanted determinism and the black box problem, philosopher and cognitive scientist Zednik investigated 'explainable AI,' and the possibility of a normative framework for it (Zednik, 2019). Zednik (2019) suggested that in order to contest the obscurity that lies within the core computations of machine learning models, and subsequently the lack of trust and understanding of their AI applications, stakeholders must be identified and efforts must be made in order to provide a more in depth understanding of the epistemically relevant elements of the AI programs as it pertains to these stakeholders. Through allowing stakeholders further insight into the ins and outs of the machine learning models, or as much as can be shown and explained per relevance to different types of stakeholders, Zednik (2019) supposes they are gaining transparency into the AI programs they are using, and thus combatting the black box problem of AI.

A few years past Zednik's normative framework initiative for XAI, von Eschenbach (2021) chimed in with further context that XAI has come to reference the models which are developed in order to respond to the black box problem. These models are considered interpretive models and are meant to shed light on different aspects of the black box system, like perhaps the decision of the machine learning model or the process or function of the system (von Eschenbach, 2021, p. 1615).

This might look like a simplified interpretation of the system resulting in something like a decision tree model. Another useful interpretive model could also be something like a heat map, where a stakeholder might resultingly be able to understand better which features were important for a system to make a classification such as from a picture (von Eschenbach, 2021, p. 1616).

An example of this could be if a social robot used in an educational setting was adapted to use a camera in order to analyze and interpret a child's perceived mood, and then respond according to their perceived mood. It may be reassuring in some way for a parent to be able to physically see a heat map indicating the areas of importance that were used by the system of the social robot in order to make an evaluation on their child's mood, to have some sense of understanding as to why it interpreted x, y, or z mood or to ensure that it is only collecting data from or focusing on pertinent areas of interest and not others.

Though understandable, people's decision to not use or to refute the inclusion of AI technologies in their lives or their children's lives due to this 'unexplainability' factor and the lack of transparency, and the accompanying lack of trust in the technology, may be, to some degree, uncalled for. Zerilli et al. (2018) point out that humans are not fully transparent either, and that it is perhaps a bit of a disproportionate response to completely dismiss the use of a tool which provides the possibility for immense good simply because people cannot fully understand it.

It is not just AI in this world which provides inadequate transparency, and for Zerilli et al. (2018), AI is argued to be held to both a higher standard and a higher expectation of transparency and understandability than even humans are, especially when making decisions, for example. Unlike AI technologies, humans are able to justify the decisions they make based sometimes purely on the feelings that they experience, and emotions both influence and act as valid reasoning for people's behaviors and decisions— no further understandable or black and white logic necessary (Zerilli et al., 2019, p. 668).

And this is of course not true for complex technologies like AI, so Zerilli et al. (2018) make the case that there is perhaps a double standard in place for those who do not trust AI simply because they cannot comprehend its decision making processes. To use the example of a doctor, someone may not be able to understand all of the biological and scientific reasoning behind a doctor's diagnosis or recommendation, but they are likely to still choose to trust in and listen to the doctor's advice (von Eschenbach, 2021, p. 1619).

This is why von Eschenbach (2021) instead stresses that trusting AI itself is maybe the wrong debate, but the question is rather whether users can trust the creators, implementers, and other users of the technology, or in other words the socio-technical system surrounding the given technology. According to von Eschenbach (2021) this is where the trust truly lies in the end anyways, and this 'web of trust' is what is necessary for trust in AI systems.

As with the case of the doctor, a patient may not understand the logic behind a decision but they do put trust into the expert who is vouching for the logic behind the decision, and they do put trust into the technicians who assist the doctor in carrying out their procedure, and they do put trust into the regulatory systems in place which certify and recertify these people as experts in their fields (von Eschenbach, 2021, p. 1619).

So, despite the black box problem within AI and the lack of true transparency, it is still of course a technology worth utilizing and trusting- responsibly. And this can be accomplished through enacting XAI measures to ensure that the machine learning, or other underlying algorithm, has some type of model which provides whichever levels of explainability necessary to satisfy stakeholders' varying needs for transparency, and also through explaining, and ensuring the upkeep of, trust in the socio-technical system.

4.2.2 AI for Social Good & the EU's ethical principles for AI

While AI as a technology does indeed present unique challenges and threats to society, many conscientious scientists, researchers, and engineers like Floridi et al. (2018, 2020) and Hager et al. (2019) have come to acknowledge and embrace this tool's powerful abilities and all the possibilities that it may offer in aiding the efforts towards a more harmonious, equitable, and prosperous global society. These proponents of AI as a tool for societal benefit have encouraged the AI4SG initiative, which seeks to leverage AI as a force for good in addressing a variety of societal challenges, as well as in improving human well being and promoting human values and interests (Floridi et al., 2018; Hager et al., 2019; Floridi et al., 2020).

This AI4SG initiative finds its foundation in the ethical design and deployment of AI and related technologies to ensure that their benefits are, amongst other things, equitably distributed, and especially that harms and potential harms are minimized. This sub section provides further insight into the principles and essential factors underpinning the AI4SG initiative, drawing on key findings primarily from the work by Floridi et al. (2018) and Floridi et al. (2020), as these principles and factors are important considerations when utilizing the VSD approach in relation to AI development. This sub section will also include a short overview of the EU's four ethical principles for AI (European Commission, 2019), as this is also of significance for Umbrello and van de Poel's (2021) updated VSD approach which is AI focused.

4.2.2.1 The five ethical principles for AI development

With the aim of developing a 'Good AI Society,' Floridi et al. (2018) as the AI4People Scientific Committee outlined five ethical principles that should be used as a guide for AI development. They began by analyzing the initiatives that were already in effect within this purview, and then offered a "synthesis of [the] existing sets of principles produced by [these] various reputable, multi-stakeholder organisations and initiatives" (Floridi et al., 2018, p. 695). This synthesis from Floridi et al. (2018) produced 47 principles overall, but when dissected more closely, the plethora of principles ultimately boiled down into five overarching principles which could be referenced for guiding the ethical development of AI.

The first four of these principles actually mirror the four core principles used notably in bioethics: "beneficence, non-maleficence, autonomy, and justice" (Floridi et al., 2018, p. 696). Floridi et al. (2018) made the point that bioethics is the most similar, in terms of the various areas of applied ethics, to digital and computing ethics. They go on to establish the correlation of these four principles in bioethics to their apropos application in addressing the ethical issues raised by the creation and use of AI technologies.

They do, however, note that these four ethical principles do not fully encompass the challenges of AI, that upon further analysis they suggest the addition of one more principle in order to complete the AI specific ethical principles (Floridi et al., 2018, p. 696). The challenge which is unique to AI was explained in Sub Section 4.2.1: how AI works is difficult to explain and this lack of transparency can weaken trust in the technology, and thus the fifth and final ethical principle for AI development is *explicability*.

Beneficence: The comparative analysis Floridi et al. (2018) conducted found and grouped certain types of phrasing like "well being," "common good," "human dignity," and "sustainability" (p. 696). All of these such phrases were used in ways which essentially conveyed the principle of beneficence, that the use of AI should be of benefit to humans, to society, and to the planet. At the heart of this principle is the understanding that AI should promote well being, empower the greatest number of people possible, preserve human dignity, enhance human capabilities, and sustain the planet and all life found in the environment.

Non-maleficence: It is not enough, when dealing with a highly impactful tool like AI, to promote the creation and use of it for only good, as is the goal of the principle of beneficence. Creators and developers of AI and similar technologies must also adhere to the guiding principle of not doing harm, or non-maleficence. Floridi et al. (2018) discuss how not only is it possible for AI to be created to do harm, but it is also possible for AI to be created for benign purposes yet still be very realistically used to do harm. They expand that this can include harm done intentionally, as in misuse of the technology, or this can be harm done accidentally, such as overuse of the technology.

Because of these types of harms, which can be either deliberate at the hands of the creators or users or as an unpredictable behavior of the system itself, it is important to consider also including some type of 'upper limit' on the technology and its capabilities, especially when it may be made available for use outside of secure and protected environments (Floridi et al., 2018, p. 697). For all of these reasons, the ethical principle of non-maleficence is also essential to AI development guidelines- AI should not cause harm

and its creation and implementation should also include safeguards in order to mitigate risks and prevent misuse.

Autonomy: The value of autonomy is made complicated by technologies like AI, since in some cases a user may, in a sense, be opting out of at least some of their autonomy by choosing to use AI and subsequently giving up some of their decision making power (Floridi et al., 2018). Floridi et al. (2018) discuss how, in this context, autonomy is linked to the value of human choice, which is why this ethical principle translates to protecting that value. And this is accomplished through adherence to the guidelines as follows, as Floridi et al. put it, "not only should the autonomy of humans be promoted, but also the autonomy of machines should be restricted and made intrinsically reversible, should human autonomy need to be re-established" (2018, p. 698).

In other words, to preserve the value of human choice, technologies like AI should be developed in such a way that the user is able to decide if and when to cede decision making power to the machine, as well as how much power to cede, and they should always have the ability to override a machine's decisions. In this way AI can still support human autonomy and decision-making, while avoiding undue manipulation or control.

Justice: This ethical principle, when synthesized between all the various sources by Floridi et al. (2018), was presented with an array of interpretations. 'Justice' in the more traditional sense was translated into the context of AI as using AI to correct or reverse wrongs of the past, as in the elimination of unfair discrimination (Floridi et al., 2018). It was also interpreted by other organizations to mean that AI should be developed and used in ways that promote fairness, as in that the benefits that AI brings should be shared amongst society and also that the technology should be made available equitably instead of as a tool for only the wealthier factions, for example (Floridi et al., 2018). And finally, Floridi et al. (2018) found that the ethical principle of justice was also accounted for through the guideline of preventing new harms from arising as would be the case if the creation of an AI application incapacitated existing social structures.

Explicability: The fifth and final ethical principle for AI development is concerned with ensuring an AI system is intelligible, that the transparency issue associated with AI's complex inner workings is in some way accounted for and the models and function of the system are understandable to some degree (Floridi et al., 2018). Through a better understanding of how the technology works, by ensuring the AI system is intelligible, it can be better deduced which good and which harms it may be enacting in society, and in which ways, which complements the ethical principles beneficence and non-maleficence (Floridi et al., 2018, p. 700).

Explicability here not only enhances trust, but it also addresses the issue of accountability. Correlating to the ethical principle of autonomy, in order to make a sound decision which maintains human autonomy, like deciding whether or not to let an AI system make a decision, a user must have some informed knowledge about how the AI system would make its decision (Floridi et al., 2018). And in order to hold the appropriate party accountable should a bad outcome arise, some understanding of why this negative outcome occurred would be necessary, thus also complementing the ethical principle of justice (Floridi et al., 2018, p. 700). So not only is explicability an ethical principle which accounts for the unique ethical issues which arise with the development of AI, but it also provides additional support for ensuring the other ethical principles are followed effectively as guidelines.

4.2.2.2 The seven essential factors of ethical AI design

In addition to Floridi et al.'s (2018) five principles for the ethical development of AI, Floridi et al. (2020) introduce seven essential factors for aligning AI design towards social good and to aid in ensuring AI technologies are ethically sound, practical, and beneficial. These seven factors are rooted in the five principles and are interdependent and vary along with each other, and therefore are not presented in or to be understood as a particularly ranked order. This sub section will provide an overview of these seven principles, which are "falsifiability and incremental deployment, safeguards against the manipulation of predictors,

receiver-contextualized intervention, receiver-contextualized explanation and transparent purposes, privacy protection and data subject consent, situational fairness, and human-friendly semanticization" (Floridi et al., 2020, p. 1775).

Falsifiability and incremental deployment: In following Popper's (1963) concept of falsifiability as a core tenant of scientific knowledge, this concept has also been translated into a core tenant of ethical design in AI technologies. According to Floridi et al. (2020), AI systems should be designed in such a way that their claims and their functions can be tested. They follow this up with the assertion that this empirical testing must be carried out continuously, that the application must be deployed incrementally and continuously assessed at each stage of this gradual introduction of the system. This rigorous testing and incremental deployment helps ensure the system is only introduced into real world environments once it has achieved a minimum safety level (Floridi et al., 2020). And as Floridi et al. (2020) point out, this stepwise process allows for ongoing monitoring as well as adjustment based on practical feedback, identifying and mitigating potential issues before they have the ability to cause significant harm.

Safeguards against the manipulation of predictors: In service of the preservation of both effectiveness and trustworthiness, Floridi et al. (2020) suggest the implementation of safeguards which protect against the manipulation of input data, to prevent biased or unreliable data from being used to train models and their predictions. These safeguards also help prevent the potential gamification of the AI system, which in the field of privacy preserving machine learning means essentially that if someone can observe the inputs of the system and the resulting outputs then they can potentially gamify the system in such a way that they manipulate the inputs in order to achieve a desired output, as touched upon by Boscoe (2019) (see also Papernot et al., 2017). This means, where applicable, knowledge of how AI system inputs affect outputs should be obscured or limited (Floridi et al., 2020, p. 1779).

Receiver-contextualized intervention: Regarding the preservation of human autonomy, Floridi et al. (2020) propose the enforcement of 'optionality,' that users must be

given options on how much and which kinds of personal information of theirs is used and therefore how personalized their experiences are with the technology. This also entails user and stakeholder input on their goals and preferences throughout the design process, as well as contextualizing in which circumstances the system performs an intervention, always allowing the user to modify their preferences of these things (Floridi et al., 2020, p. 1780).

Receiver-contextualized explanation and transparent purposes: Also in the efforts of transparency and combating the black box nature of AI systems, Floridi et al. (2020) suggest that, by default, users of the systems and the receivers of the output of the systems be provided appropriate explanation of and argumentation for the system's objectives. This should also include understandable and relevant explanations for the system's decisions (Floridi et al., 2020, p. 1784).

Privacy protection and data subject consent: As has been established, privacy protection is a major and elemental requirement for the ethical design of AI technologies. Floridi et al. (2020) stress that this privacy encompasses of course personal data collection and uses, and that this also includes informed consent about which data may be collected, as well as how it is collected, and how and why it is used. In toto, "designers should respect the threshold of consent established for the processing of datasets of personal data" (Floridi et al., 2020, p. 1786).

Situational fairness: Not only is it important for the design of AI systems to include informed consent about the data they collect and use, but Floridi et al. (2020) assert that designers should also ensure that the training data they use is explicitly clear of any kind of bias, in order to avoid biased results in significant outputs like those in decision-making uses. They specify that this situational fairness also entails removing irrelevant variables from datasets used for training the systems, however, that this is not necessarily the case when the inclusion of these irrelevant variables may support ethical values like safety or inclusion (Floridi et al., 2020, p. 1788).

Human-friendly semanticization: The final factor which is essential to the ethical design of AI technologies is centered again on preserving human autonomy, and also promoting human values and interests. According to Floridi et al. (2020), "AI should be deployed to *facilitate* human-friendly semantisation, but not to provide it itself" (p. 1789). In other words, AI systems can and should be made and used to help humans make sense of things in ways that are understandable to them, but the systems should not, however, be created to ascribe meaning to things of their own accord. As far as design guidelines go, Floridi et al. (2020) declare that in order to maintain AI4SG ideologies for a Good AI Society, designers should ensure their creations do not hinder human ability to make meaning.

4.2.2.3 The EU's four ethical principles for AI

Just to provide a quick overview of this, the European Commission formed an EU High-Level Expert Group on AI which put together a report in 2019 outlining their four ethical principles for developing and deploying AI systems. Each of the seven essential factors detailed in the previous sub section relate in some way and can be mapped onto these four ethical principles. This mapping is important since any deviation from these four more generalized values has the potential for harmful consequences. The EU Commission's four ethical principles are not so different from the five ethical principles discussed in Sub Section 4.2.2.1 which were laid out by Floridi et al. (2018), however, there are some subtle additions that will be highlighted in this small sub section.

The "EU High-Level Expert Group on AI" outlines their chief ethical principles as follows: "respect for human autonomy, prevention of harm, fairness, and explicability*"* (Umbrello & van de Poel, 2021, p. 286). The first, ***respect for human autonomy***, is again focused on emphasizing that AI systems should enhance human decision making rather than diminish or replace it, however, the European Commission (2019) also includes that an AI system should not coerce or manipulate people into making decisions, especially those that may be against their interests or that might disregard their informed consent. This

expansion to the principle of autonomy in AI development is particularly pertinent when considering the use of AI for moral enhancement, and especially so when considering this type of use with regard to children in early education who are highly impressionable and swayed by even suggestion. Even the possibility of coercion and manipulation must be avoided at all costs in any type of AI application of this nature.

The principle of ***prevention of harm*** is similar to the aforementioned non-maleficence, that when designing AI systems identifying and mitigating risks and incorporating fail-safes are necessary to avoid causing harm. In this principle however, the European Commission (2019) specifies that this harm encompasses both physical and psychological well being, and this is a distinction worth noting with regard to the possible use of social robots which are meant to aid in moral education in early education. Children using this type of tool must be safe from not only adverse psychological effects, but also any possible physical harm that may occur due to the robot and its moving parts.

The third principle, ***fairness***, is concerned with avoiding bias in an AI system's operations, as well as avoiding discrimination and any kind of promotion of inequality. Here, though, the European Commission (2019) aims at the prevention of marginalization of any group of people through the emphasis of the consideration of diverse needs and perspectives accomplished by ensuring inclusive design processes. This is of course again applicable when concerned with the development of a social robotics tool meant to be used in early education and meant to help teach children about morality and build their moral reasoning skills, including diverse stakeholder inputs in the design process will help ensure fairness is maintained in the design process.

And finally, ***explicability*** is much the same as was outlined in the earlier sub section, ensuring an AI system is understandable and transparent to as high a degree as possible is necessary to build trust in the technology and assure accountability. The European Commission (2019) mentions also that the transparency involved here also includes providing clear and accessible information and explanation about the data that is used and the decisions made in addition to how the system works. These factors are important of

course in particular to stakeholders like parents in the case of the use of social robots in early education.

4.2.3 Combining AI ethics principles and the VSD approach

With a more solid understanding of one of the foundational challenges presented by the use of AI technologies (the black box problem, as outline in Sub Section 4.2.1), and along with a better understanding of the efforts encompassed within initiatives like AI4SG which are meant to mitigate this and the other challenges AI presents (the collective ethical principles of AI development and the essential factors of AI design, as outlined in Sub Section 4.2.2), it can be better understood how the Value Sensitive Design framework can be adapted to accommodate the unique considerations necessary for ethical AI design, development, and implementation. Umbrello and van de Poel (2021) propose the modification of the VSD approach in the following three ways:

> (1) integrating AI4SG principles into VSD as design norms from which more specific design requirements can be derived; (2) distinguishing between values promoted by design and values respected by design to ensure the resulting outcome does not simply avoid harm but also contributes to doing good, and (3) extending the VSD process to encompass the whole life cycle of an AI technology to be able to monitor unintended value consequences and redesign the technology as needed. (p. 288)

4.2.3.1 Integrating AI4SG principles as norms

The first modification (1) is understood better through the lens of the value hierarchy- that values comprise norms, and from those norms designers can more easily formulate design requirements. This was largely accomplished in Sub Section 4.2.2— the collective principles outlined by Floridi et al. (2018) and the European Commission (2019) serve as the overarching values to be considered in AI development, and the seven essential factors from Floridi et al. (2020) comprise the norms which inform the more specific design requirements associated with the various values. Umbrello and van de Poel (2021) use the EU's four ethical principles onto which they map the seven essential factors as follows:

- the value ***'respect for human autonomy'*** finds its norms as the three essential factors *receiver-contextualized intervention, privacy protection and data subject consent,* and *human-friendly semanticization*;
- the value ***'prevention of harm'*** encompasses the essential factor *falsifiability and incremental deployment*, as well as again *privacy protection and data subject consent*;
- the value of ***'fairness'*** is composed of the norms of *safeguards against the manipulations of predictors,* and *situational fairness*;
- and lastly, the value of ***'explicability'*** is understood better through the essential factors of both *receiver-contextualized explanation and transparent purposes,* and again *human-friendly semanticization*.

As mentioned, the seven essential factors function as norms associated with the ethical principles as values, and through these factors more specific design requirements can be established as outlined in Sub Section 4.2.2. This modification to the VSD approach elaborates on van de Poel's (2013) work of translating values into design requirements, a tricky but important step to methodologically implementing this in practice. This first modification also provides a solution to one of the criticisms and challenges associated with the VSD framework, as discussed within Section 4.1, that in the traditional VSD approach there is no prescribed way to transcribe values into design requirements.

4.2.3.2 Distinguishing values

The second modification (2) to the VSD approach from Umbrello and van de Poel (2021) proposes more emphasized focus on contributing to social good, that in order to utilize AI technologies in ways that go beyond simply avoiding harm, the AI4SG goals can be aligned with the Sustainable Development Goals (SDGs) put forth by the UN (Schwan, 2019). This orientation towards more actively promoting socially desirable outcomes would ensure that AI technologies are employed in efforts like ending poverty and hunger (SDGs 1 and 2), quality education and decent work and economic growth (SDGs 4 and 8), and reduced inequalities and peace, justice, and strong institutions (SDGs 10 and 16), as outlined by

Schwan (2019). Incorporating the SDGs into the goals and values of AI4SG and the VSD approach for AI, this field can contribute more towards SDG 17, partnerships for the goals.

In the interest of this combining of global efforts, I would like to put forward the additional inclusion of the NAE Grand Challenges for Engineering (2008), whose goals largely overlap with the SDGs, but take them on from an engineering perspective. This distinction can be particularly prudent for supplementing the AI4SG and VSD for AI goals, as many of the Grand Engineering Challenges (GECs) can be significantly aided through the use of AI. The subsect of these types of challenges are as follows: advance health informatics; reverse-engineer the brain; secure cyberspace; enhance virtual reality; engineer the tools for scientific discovery; and, of particular importance here, advance personalized learning (NAE Grand Challenges for Engineering Committee, 2008).

These such challenges are especially topical when it comes to harnessing AI technologies ethically and responsibly, and, along with the SDGs, the focused efforts towards these collective goals can prove instrumental in transforming society towards more positive outcomes. In this way, Umbrello and van de Poel (2021) suggest that distinguishing between values that are "promoted" by the design and those that are "respected" by it may encourage the use of AI towards more active goals instead of merely the passive goal of avoiding harm. And, through introducing the GECs here I aim to fill the gap in this field which has, from my findings, not included these Grand Engineering Challenges, despite the incredible applicability to the field and the immense good that can come from highlighting this effort and its goals.

4.2.3.3 Extending VSD

Due to the nature of AI and ML technologies, Umbrello and van de Poel (2021) emphasize as their third and final modification to the VSD approach (3) that designers *must* extend the VSD approach to include the *entire* life cycle of the technology, from conception through deployment- and beyond. AI systems that learn and develop as a byproduct of their use pose significant potentials for harm, and thus increase the pressing need for these

technologies to be incrementally deployed and continuously assessed, tested, and adapted when straying from the intended value alignments.

4.2.3.4 The four iterative phases of adapted VSD

The preceding three modifications to the VSD approach are proposed from Umbrello and van de Poel (2021) to be enacted in the following four iterative phases: (i) context analysis, (ii) value identification, (iii) formulating design requirements, and (iv) prototyping. Context analysis (i) entails an investigation into the context of use for the technology, as well as utilizing empirical investigations in the analysis of the values that arise from a more clear understanding of the motivations for the design, the sociocultural and political environment, and the stakeholders involved (Umbrello & van de Poel, 2021, p. 289). Context analysis according to Umbrello and van de Poel (2021) ensures that the values considered during the design process are relevant and meaningful within the specific setting of the AI application.

The second phase is accomplished through both the empirical and conceptual investigations found in the traditional VSD approach. For Umbrello and van de Poel (2021), value identification (ii) simply means identifying a set of values which should guide the design process. These values are derived from three sources: values promoted by the design (e.g., those aligned with the SDGs and/or GECs), values that need to be respected (e.g., AI-specific values such as fairness and explicability), and context-specific values that emerge from the contextual analysis (Umbrello & van de Poel, 2021, p. 289). Identifying these values aids in ensuring these values are both normatively sound and practically applicable within the given context.

Formulating design requirements (iii) is guided by both the values identified in the second phase (ii) and the contextual analysis performed in phase one (i) (Umbrello & van de Poel, 2021, p. 289). Using the value hierarchy discussed in Sub Section 4.2.3.1 is particularly helpful here to translate values into design requirements. Umbrello and van de Poel (2021) suggest that promoted values (e.g., SDGs and GECs) should be translated into

criteria that the design should strive to meet, while respected values (e.g., fairness and autonomy) should be translated into boundary conditions that the design must not violate. Context-specific values function most likely in shaping how these design requirements are formulated, ensuring that the design is appropriately tailored to its intended environment per stakeholder input (Umbrello & van de Poel, 2021, p. 290).

The final phase (iv) of Umbrello and van de Poel's (2021) adapted VSD approach for AI involves the creation of tests and prototypes that meet the established design requirements. This phase is not a one-time effort but is instead extended across the entire life cycle of the AI technology, since, as the technology is deployed and evolves, it may develop in unexpected ways, requiring further iterations of the design process (Umbrello & van de Poel, 2021, p. 290). The suggested continuous monitoring and redesign put forth by Umbrello and van de Poel (2021) ensure that the AI system remains aligned with its intended values and can adapt to any new ethical challenges that arise.

In sum, by integrating AI4SG principles as norms which inform value-based design requirements, distinguishing promoted, respected, and contextual values, and extending VSD to the entire life cycle, Umbrello and van de Poel's (2021) adapted VSD for AI approach aims to address the unique challenges presented by AI technology design and also to more actively encourage the use of AI to serve in the efforts towards social good.

4.2.3.5 VSD and social robots

The Schmiedel et al. (2022) research group carried out several VSD projects which largely centered around AI technologies, and more specifically social robots. As such, they synthesized some best practices based on their experiences and put forth suggestions for the field moving forward. I will quickly recap their findings, as they proved useful in guiding me in the process of this project. First, they reported a review which revealed most VSD projects do not iterate their processes, and subsequently they suggest a multi-iterative approach which is made up of (i) value identification, (ii) value embedding, and (iii) value evaluation (Schmiedel et al., 2022, p. 76).

Value identification (i) encompasses, according to Schmiedel et al. (2022), all three phases of the traditional VSD approach: *conceptual*, values can be identified through the relevant literature; *empirical*, through reported stakeholder context-specific values; and *technically*, as a technical artifact they found mock-ups particularly useful when researching social robots as most people have not interacted with them and it is its own endeavor to actually create and implement them in full. Value embedding (ii), which mimics Umbrello and van de Poel's (2021) translation of values into design requirements, and adds that this phase in social robot design may occur iteratively in and of itself between "technical prototyping and empirical usability testing" (Schmiedel et al., 2022, p. 76). And value evaluation (iii), which Schmiedel et al. (2022) assert empirically assesses the use of the social robot for the correct appearance of intended values and any unanticipated values or new value tensions which may arise— advising the return to value identification to ensure the design remains in scope of the stakeholder values.

Their second advice was to start with a specific use case, since they found that not all values are universally held through each specific use of social robots and that a context-specific approach proved more useful especially when considering stakeholder's prioritization of values was dependent on use case bases (Schmiedel et al., 2022, p. 77). Schmiedel et al.'s (2022) final suggestion was that those in the field combine efforts towards a VSD catalog which is technology-specific, as this will foster more effective exchanges in the community and provide best practices to leverage in the furthered development of value sensitive technologies.

4.3 VSD applied to social robots in early education

With a more foundational understanding of VSD and its applicability for AI and social robots, we can more readily examine relevant value considerations in the design of a social robot system intended to aid in moral education in early educational practices. Following this VSD approach, these identified values can subsequently be assessed for how they may be translated into design requirements for a technological solution in this context.

Although they did not explicitly incorporate the suggestions for adapting VSD to AI and related technologies, Smakman and Konijn (2019) did a systematic literature review based in the VSD approach identifying the moral values being considered in relation to the general use of robots as aids for education. Though this is not specific to the application of such a technology used for the purposes of *moral* education in early education, these identified values are still highly relevant to such an application and should be evaluated as such.

They found that in regard to 'robot tutors,' the literature reported values like psychological welfare and happiness, freedom from bias, efficiency, and usability to be *both* positively and negatively correlated (Smakman & Konijn, 2019). This signaled that there are a variety of possible benefits considered in the use of robots in educational settings, while at the same time there are major ethical concerns to be addressed in the design and deployment of such tools in these contexts. Other unique values they reported include human contact and friendship and attachment, along with more standard concerns like security, privacy, accountability, and trust (Smakman & Konijn, 2019).

The value concerns surrounding human contact and friendship and attachment could be interpreted in the project level guidelines as design requirements involving limited time and use of the social robot system during the educational day. This would mitigate the possibility that students form 'too much' attachment to the bots and become over reliant on them for companionship outside of limited educational purposes. From a curriculum content and implementation guideline perspective, a resultant design requirement could be that the curriculum itself is also limited to very specific and short use cases. This would more practically mean that the 'lessons' and their respective activities are small and succinct, and stick to very particular topics and activity types. Though not a guarantee that students would not become overly attached to the bots, design requirements such as these would be a beneficial step in addressing these types of value concerns.

The values Smakman and Konijn (2019) established from the literature were all considered from the viewpoints of children and teachers, noting that parents were overlooked stakeholders in much of the literature at the time. Their review concluded that the use of robots in education is morally justifiable, but that key stakeholders like parents must be referred to in the design process for more appropriate levels of ethical soundness (Smakman & Konijn, 2019).

In filling the gap of the missing parent stakeholder considerations, Smakman et al. (2020) carried out a study based in the VSD approach in order to ascertain the moral conceptions parents held regarding the use of social robots in primary schools. They cite the opportunities of the use of social robots in education as including facilitating and promoting cognitive gains particularly in children, the possibility of personalized learning, and increased incentive and enjoyment in educational activities (Smakman et al., 2020, p. 7946). By targeting specifically parents' concerns they hope the consideration of these values will give rise to wider social acceptance and adoption through the more ethical creation and implementation of this technology for this context (Smakman et al., 2020).

They collected empirical data from focus group sessions with parents in The Netherlands, resulting in derived key moral value considerations which they suggest to be translated into guidelines for the robotic industry to incorporate into a more ethical design and implementation of social robots for early education (Smakman et al., 2020). These moral values included many of the same as were identified in the literature review, however, they found that these parents presented three unique values: flexibility, responsibility, and dependability (Smakman et al., 2020, p. 7949).

Flexibility in this context, from the Dutch parents' perspective as reported by Smakman et al. (2020), meant that the bots had the possibility of providing flexibility for families to go on holiday more easily since the social robot could be brought along with them and provide the education while they were away. Though, they reported that parents also considered that the social robots may hinder flexibility if they were too large or heavy since parents may not want to bike the bots back and forth from school with their children

(Smakman et al., 2020). As a result of this identified value of flexibility, a design requirement for a system such as this would be the physical size and weight of the system being as small and light as possible- and this has the potential to align with some values from the SDGs and GECs (as discussed in Sub Section 4.2.3.2) concerning sustainability.

The identified values of responsibility and dependability were relayed by Smakman et al. (2020) through the lens of parenting responsibility, that the parents were concerned with overreliance on the technology in taking over parenting tasks, especially if the bot was meant to be taken home and support their children's development in the home as well as the school. The resulting project level design requirement of this value consideration is that parents whose children are using social robots in school ought to be informed of 'best practices' when the bot is at home, such as limiting the time that a bot is used or whenever possible to use the bot with the child. This would help the parent remain apprised of the goings on in the curriculum and also allow for more prompt stakeholder feedback if the parent notices something 'off' about the bot's behavior or content.

Under the value of efficiency, Smakman et al. (2020) describe concerns parents expressed with regard to the bots possibly giving children erroneous information, as well as the possibility that the robot could be hacked. Similar to this, parental concerns surrounding the values of privacy, security, safety and accountability stemmed largely from the unknowns with regard to the lack of policies and regulations surrounding things like data collection methods and storage (Smakman et al., 2020, p. 7951). As a result of this finding, future studies should include in their investigations research into any developing policies in this field, relay these findings in the stakeholder engagement activities, and gather any subsequent values of concern which could be translated into design requirements.

To account for the value concerns of privacy and security, something which has scarcely been seen in the literature of this field is a call for privacy preserving machine learning considerations as seen in deep learning models. This value consideration could result in a specific design requirement like the potential use of Fully Homomorphic

Encryption (FHE) (Zhang et al., 2021). Using FHE for example would essentially mean that data could be collected and immediately encrypted and stored, and then the retrieval and evaluation of that sensitive data could occur so that the system can learn from it- while remaining encrypted throughout all steps of engagement with that data, thus ensuring privacy and security are maintained even while the system accesses and evaluates the data (Marcolla, 2022).

Consideration of these values and their related concerns also reinforces the need for design requirements surrounding algorithmic and implementational transparency, such as employing XAI models (as described in Sub Section 4.2.1). In addition to this, and something else I have yet to see suggested in the literature, is that these kinds of projects ought to consider becoming an open-source entity if they are not already (or at least embracing open-source ideologies) (Gacek & Arief, 2004). Open source essentially means that anyone can view, use, and alter (for their own private purposes) the software in question, and this open collaboration results in public trust in the software.

And yes, this open ideology can be adopted while still ensuring no changes occur to the actual codebase used in the social robots. It simply means that others can take the code used in the application and enhance it in different ways that suit their unique needs on their own personal instances of the software. This type of design feature is also one way to address the value of usability, in which parents expressed concern about the lack of universal access and the obvious privilege gap associated with incorporating advanced technologies in education systems (Smakman et al., 2020, p. 7950). While they do not solve all issues, the design requirement options laid out here would be pertinent also for addressing the value concerns requiring bolstering trust from all stakeholders and heightening security for the users.

Parents and other key stakeholders also expressed worry that using social robots in education could compromise children's social and emotional development, which is a major consideration correlated with the identified value of ‘psychological welfare and happiness’ in the design of a technical application such as this (Smakman et al., 2020;

Smakman & Konijn, 2019). However, according to a study from Smakman et al. (2022), experienced teachers assert that this actually is not the case, and that in fact the use of social robots in education has reportedly even enhanced, for example, children with special needs' ability to connect with their human peers and teachers.

Although the use of social robots in educational settings seems to provide many positives, Smakman et al. (2022) also echo that the use of social robots should not supplant the role of human teachers, but rather supplement the guidance that children receive throughout their formal education. From these value considerations, translated design requirements could include on the project level that the educators using social robots in educational settings limit the use of the social robots in the classroom, ensuring that bots do not replace human interaction but simply compliment it.

Because these accumulated VSD empirical investigation findings do not reflect values resulting specifically from the context of use of social robots for *moral* education, further empirical investigations are needed in order to fill that gap. Identifying stakeholder values concerning this particular context of use would be valuable for informing further design requirements for this specialized technological solution. Despite this lapse in specific data, and instead of speculating about its possible contents, the following solution proposal will proceed as informed by the legitimate data and findings as is.

5 Proposal for an Enhanced Solution: 'MiruBots'

Individually, each of the previously outlined uses of technology from Chapter 3 (the PopBots case study and the suggestions for AI for moral enhancement) have justified reasons for implementation within society, along with complementary goals, however, neither approach of their own accord offers one unified suggested approach which specifically aims to facilitate an *enhanced moral* education within *early* education *through* the

use of technology. Therefore, I will put forth a proposed solution which combines the aforementioned approaches in a way that ensures technomoral resilience as the goal of the modern moral education curriculum, while employing techniques which will promote the effective cognitive development in the intended users, young children.

Put more concretely, in this chapter I will offer an enhanced solution which combines the effective and essential elements of each of the previous sections and chapters: the key factors and values outlined both in early cognitive development and in enhanced moral development from Chapter 2, the effective aspects of both the PopBots system and the suggested implementations of AI for moral enhancement from Chapter 3, and the values both considered and discovered in the VSD investigations from Chapter 4. I propose to call this suggested enhanced solution 'MiruBots'.[1]

This suggested solution will be composed of proffered materials and implementation guidelines, followed by a simple example use case to paint the picture of the proposal more clearly, and then an evaluation of its effectiveness at addressing the key cognitive and moral developmental factors and values identified in Chapter 2, as well as the values addressed from the VSD studies discussed in Chapter 4. This proposed solution aims to fill the gap in current literature which fails to discuss and promote a technomoral resilience focused moral education specifically in early education with the thoughtful utilization of AI-enhanced social robotics technology, while reinforcing the responsible deployment of this enhanced solution proposal through a VSD analysis of this technology for this purpose.

[1] The word 'miru' has many meanings from several different cultures. The inspiration for this application of the word comes from a few select interpretations: in Japanese 'miru' means 'to look' or 'to see,' reminiscent of the English 'mirror' meaning 'to reflect' or 'to imitate'; in Korean 'miru' can mean 'poplar tree,' which in the Celtic culture represents "transformation and vision" and whose spirit teaches us to "keep our roots strong" and helps us to "overcome…self doubts that may block our endeavors"; in Latin 'miru' means 'wonder' and 'marvel'; and in Proto-Slavic languages 'miru' means both 'world' and 'peace' (Choi, 2013, para. 4-10). Altogether these meanings express the essence and goals of the MiruBots system as a tool which helps kids to critically reflect and build moral resilience.

5.1 Suggested implementation guidelines

In practicality, MiruBots serve as one way in which young children may more effectively learn how to be a moral person in formal education settings through the use of social robots. These social robots would be equipped with AI applications which are carefully created in order to foster their cognitive and moral development through interactive and individualized moral education curricula, enhanced to include a focus on building up the capacities necessary for cultivating self reflection, emotional regulation, and open-mindedness. This section will provide an overview of suggested guidelines for both the project as a whole and the implementation and content of suggested curricula.

5.1.1 Overarching project guidelines

The MiruBots project would need to consist of an interdisciplinary team, encompassing researchers and engineers (including privacy and security oriented experts), early development psychologists, ethicists, and, where applicable, children's advocacy agents (from groups like the Children's Defense Fund (2023)). This team would inherently welcome the addition of any other kind of expert found to be relevant to the endeavor. This collection of individuals would be able to address both technical and ethical concerns, ensure the educational efficacy of the project, and effectively monitor its use and development. This project level design requirement would address many of the value concerns laid out by multiple stakeholder groups as discussed in Section 4.3, such as psychological welfare, efficiency, and usability (Smakman & Konijn, 2019).

The project as an entity itself should consider some level of being open sourced and including XAI practices. What this could more practically mean here would be something like maintaining an accessible platform, such as a page on a website, which fully discloses its methods of operation. This would include justification for which kinds of data collections may be necessary and how they may be collected, as well as justification for which algorithmic (and otherwise) technologies may be used.

In addition to justification, the site could also offer some type of understandable report describing any AI/ML algorithms which it may be using which could pose some threat to the privacy and security of the user(s). These efforts could also include a repository on GitHub (n.d.) which allows anyone in the community to access, alter, and use the code for their own purposes (without altering or pushing any changes to the MiruBots system itself). These types of project level design requirements can offer parents and other stakeholders at least some understanding of how the software being used is working. They also provide solutions to key value considerations like explicability, trust, transparency, usability, and accessibility (as discussed in Section 4.3), and the essential factor and norm 'receiver-contextualized explanation and transparent purposes' as discussed in Sub Section 4.2.2 (Floridi et al., 2020).

The research itself ought to be carried out thoroughly and include high stakeholder engagement regularly throughout design, development, deployment, and redesigns, as is encouraged through the VSD approach. Surveys, interviews, and focus groups should be formed and enacted from diverse stakeholder groups at every new addition to the curriculum, every new use case (e.g., home vs school), and every new class/year. In the survey and interview phase, it would be prudent per XAI recommendations to ask stakeholders explicitly to what level they wish to understand the underlying programming models being used in the MiruBots system, in case any new values arise which are not already included and addressed on the XAI webpage. This would garner substantial and diversified data collection, providing more robust value considerations and design requirements for a more ethically sound and effective product.

The design process in its entirety must maintain an iterative model of deployment and testing per the ethical principles discussed in Sub Section 4.2.2, and incorporate the feedback from the stakeholder engagement activities at each step (also remembering to do value evaluations to ensure values identified early on are not lost or forgotten). This provides quality assurance from both the technical standpoint and the content standpoint. Additionally, updates to the curriculum would more readily account for updates to moral

norms per the understandings of technomoral change under technomoral resilience, as discussed in Sub Section 2.2.2. In other words, moral norms understandably change and adapt as society changes and new technologies develop, impacting societal norms and values therein; thus, regular iterations of stakeholder engagement activities ensure that these fluctuating norms and values are always accounted for and incorporated in the design of the technology.

In response to the value of flexibility, as reported by Smakman et al. (2020), another project level design suggestion is to keep the MiruBots system as lightweight as possible, both in physical weight and also resources necessary. This would additionally align with AI4SG goals like sustainability and also help to address the value of usability and potentially the privilege gap associated with the limitations of access to resources like a MiruBots system.

To address the values of concern identified in Section 4.3 surrounding attachment, friendship, and human contact (Smakman & Konijn, 2019), a project level design requirement would be to ensure that the educators responsible for using this bot system in the classroom understand it as a supplement to their teaching and use it sparingly, instead of relying on it to occupy students when they are overwhelmed or if the student is bored, for example. Furthermore, a cumulation of 'best practices' should be made available to educators and parents which outline guidelines for use of the bot, such as time limits and whenever possible using the bot with the parent. This design suggestion would address the concerns associated with the values of responsibility and dependability as was also discussed in Section 4.3.

A final recommendation for the MiruBots project would be to include, however possible, initiatives which encourage the expansion and development of projects of this nature especially in underserved areas. This includes initiatives which distribute and promote educational material to educators to enable them to better educate their students on AI, AI ethics, and diverse morality philosophies and curricula, even when they cannot use the physical MiruBots system itself. This would align also with goal 17 of the SDGs as

outlined in Sub Section 4.2.3.2, “partnership in the efforts towards Social Good” (Schwan, 2019). This alignment is in conjunction with Umbrello and van de Poel’s (2021) suggested adaptation of the VSD approach for AI and social robots, such that this project level design requirement results from distinguished values which intend to promote the creation of the technology explicitly for good, rather than creating a technology which merely passively avoids doing harm.

Although these overarching project guidelines are based on the results of the investigations into the various value considerations for this type of project, these design suggestions are incomplete. Further project level design requirements for this use of social robots could be made more explicit after extending these investigations to include research into current and future policies for AI and social robots in educational spaces around the world, as well as further examination of moral education curricula and policies in various educational systems across the globe.

5.1.2 Curriculum content & implementation guidelines

The physical set up of this proposed MiruBots system would be largely based on the existing PopBots platform: a smart phone for the emotive social robot’s face, a variety of LEGO blocks for the child to build the body how they wish, motors and sensors for the child and parent or guardian to implement any moving features on the social robot, and a tablet for the child to visually and tactilely interact with the programmed curriculum (Williams et al., 2019).

As a general recommendation, though the interactive and ‘fun’ nature of the social robot system is motivating and engaging for children as Williams et al. (2019) point out, this does not mean that the visual contents of the system need to be overly stimulating. As Javed et al. (2019) discuss in their research, limiting stimuli when using social robots with young children may be preferable for many reasons, especially when considering the negative effects on children with sensory processing issues. In the interest of non-maleficence, the recommendation here is to adopt low-stimulation strategies in

content delivery methodologies, and this is supported also by child psychologists Rodrigues & Pandeirada (2018).

Low stimulation here could mean that the tablet platform interface uses non-vibrant colors with very few objects to interact with on the screen, as well as limiting any sound output to only that which is necessary. Further research into this could produce more specific design requirements of this nature such as more appropriate color palettes to stick to and types of audio frequencies more suited to an application of this nature with this goal. This design requirement also addresses some value concerns like 'psychological welfare' and usability, as discussed in Section 4.3 (Smakman & Konijn, 2019).

The content of the application would range, with a variety of modes which reflect different educational goals including, but not limited to, AI concepts, AI ethics, building moral reasoning skills, and building emotional regulation skills. A core tenant within any such curriculum implementations would strictly entail that the system never prescribes meaning or ethical importance of its own accord, preserving human centered semanticism per the ethical principles and norms of the adapted VSD approach as outlined in Sub Section 4.2.3.

This could be enacted through pre-selecting moral dilemma situations for consideration within the curriculum, contrary to allowing the system to come up with its own situations for the children to discuss and work through. This would mitigate unpredictable behaviors from the bot and therefore mitigate unexpected or unwanted impacts on the child, further protecting the value of 'psychological welfare'. Having a white list instead of a black list (as in, selecting which things to allow as opposed to which things to *not* allow) also ensures that human values are maintained.

Though on the surface this runs the risk of being seen as a dictatorial design suggestion, when thoughtfully enacted in alignment with ethical principles (such as those outlined in Sub Sections 4.2.2 and 4.2.3), the realistic result of this design requirement is that children would only be exposed to ideas which are pre-determined and pre-approved through stakeholder and expert engagements. The system would therefore have some

safeguards in place also against the manipulation of predictors, successfully addressing values like fairness, autonomy, security, and safety.

Similarly, the system should never assert what is right or wrong, its function in this regard must simply be engaging the child in ways which encourage the development of their critical reasoning skills, allowing them to reflect on their own assumptions and beliefs in a safe and structured environment. Informed by the findings from Section 3.2, much of the content would be enacted in the form of a socratic dialogue partner- never declaring any one choice as morally superior, but rather introducing questions which allow the child to consider and weigh the consequences of a given action of their own accord. This aligns also with the moral development value of building the capacities for technomoral resilience as discussed in Sub Section 2.2.2, and the cognitive development value of the dialectic method as discussed in Sub Section 2.1.1.

Stakeholders pointed out an additional benefit to using a social robot for this type of classroom interaction is that the bot will remain objective whereas parents and teachers may not, and this alters a student's experience with learning the content (Smakman et al., 2020). This suggestion is also supported by previously referenced education and development researcher, Chazan, who said that "the teacher's role is to explicate, not propagate views, [involving] an ability to utilize and model the Socratic method of questioning, a sensitivity to group dynamics, and the ability to summarize without preaching" (Chazan, 2022, p. 30). Social robots have the power to enact all of these abilities in educational settings, when implemented thoughtfully.

Some of the modes may enact this socratic and dialectic method through different ethical frameworks, as suggested by the investigation results from Section 3.2. Though the very young kids may not need to know what precisely a 'deontological' viewpoint is, these types of signifiers may be useful to introduce as the curriculum progresses throughout the age groups. Regardless, having a mode which presents a moral situation and then discusses it from varying viewpoints can help develop the child's moral imagination per the technomoral resilience guidelines discussed in Sub Section 2.2.2.

Another mode of operation may include a group mode, which could present scenarios for consideration to more than one student and facilitate peer discussion, an essential element of moral development and reasoning skills (as discussed in Section 2.2). This mode would also encourage human to human contact and interactions (addressing value concerns of human contact and usability as discussed in Section 4.3), and offer more tangible experiences for children to build the capacity for maintaining their moral agency (addressing the value of technomoral resilience as discussed in Sub Section 2.2.2).

Since they are likely to engage in more emotionally escalating dialogue with other children when discussing differing viewpoints, as opposed to with the MiruBots system alone, this mode would allow children opportunities to practice their emotional regulation skills along with their moral reasoning skills per technomoral resilience guildelines. This group mode design suggestion also aligns with the values of fairness and usability, perhaps also addressing sustainability and/or the privilege gap given that one bot could be used for a whole classroom.

The curriculum should also incorporate elements of healthy representation and mimicry, in the sense that the social robots demonstrate things like good manners (e.g., saying please and thank you), consideration of others (e.g., encouraging inclusion of others in discussions), and even good emotional regulation skills (e.g., taking a deep breath when discussing difficult topics). As established in Chapter 2, observation and imitation are key developmental properties, so social robots present a unique opportunity to provide children with representations of specially curated behaviors which are deemed healthy and which stakeholders and experts would like to encourage in the upcoming generations.

Along these lines, another element of the curriculum should include emotional regulation practices, as is encouraged under the guidelines of building technomoral resilience. This can be its own lesson topic, and it could also be included optionally as a mode in which emotional regulation practice is intermittently interjected throughout the time spent in other modes while focusing on other educational topics. Speaking of, optionality should be implemented in every way possible in order to address the norm of

'receiver-contextualized intervention' and the value of respecting human autonomy. If there are different modes which require different levels of user input, for example, that encompass sensitive data collection of any kind or if these modes differ in intervention frequency or potency then it should be up to the children, the parents, and the educators as to which and how much of these things are affecting the user and their experience.

With respect to the concerns surrounding the stakeholder identified value of attachment, the curriculum design should also ensure that lessons and activities are 'bite-sized,' so to speak. In order to lessen the potential for unhealthy attachment to the social robots, the content should be limited to concise lessons which, for example, effectively explain a moral concept and have an associated interactive activity to demonstrate said concept and allow the child to engage with it, and then the lesson and the interaction ends in a timely manner. The curriculum should be structured in such a way that no lesson takes 'too much' time, and the content should be broken down into succinct correlated activities.

While the curriculum content and implementation design requirements suggested here are a result of the investigations carried out which identified key values in cognitive and moral development, technical design, and stakeholder values, these suggestions are not exhaustive and could be further improved and tailored. These design requirements do not fully encompass all of the ethical considerations necessary for an actual implementation of this nature, and should be bolstered by more in depth research into areas like age appropriate educational content and effective moral educational activities. Such further studies should also include data collection from stakeholders specifically concerning moral education as it relates to young students and the use of social robots.

5.2 Painting the picture with a simple use case

As suggested under the guidelines for developing technomoral resilience, one (in this case, a young student) should immersively practice engaging with, ideally, both real and imagined moral situations which are presented due to technological impact on society. So,

suppose a child's class is privileged enough to have at least one MiruBots system, and suppose it is this child's turn to use the system. When they get to the bot they (or the teacher) turns on the system and logs in using their personalized pin to the child's specific profile which contains their preferences and curriculum progress.

The MiruBot welcomes the child and asks how they are doing today. The child responds that they are excited to play with the bot today, so the bot goes on to ask the child which lesson they would like to go through today. The child uses the tablet to choose from the available lessons they have left to complete for the unit they are currently in.

The child chooses 'Consent and Privacy,' so the MiruBots system presents a scenario to the child on the tablet while the bot narrates: "Sam takes a picture of their friend Alex during lunch. Sam wants to share the picture with all their friends online." When the narration finishes, the bot blinks and says, "Hmm.. What do you think could be wrong in this situation?" The child responds with their thoughts. The MiruBots system asks the child, "Should Sam ask Alex for permission before posting the picture?" The child responds with their thoughts.

The bot follows up with, "This is called asking for consent, when you check with somebody whether or not they are okay with you doing something involving them. Have you heard of consent?" The child reflects and responds. The bot continues, "Asking permission before posting pictures of other people online is important, so you can respect their privacy *and* their feelings. What feelings might you experience if someone does something like this without your permission?" The child again engages their imagination, reflects on their experiences and feelings, and responds.

The bot asks a final question, "What might you have done differently in this situation?" The child thinks and responds, and the MiruBots system finishes the lesson with, "I may not have taken the picture of Alex in the first place, not without getting Alex's consent anyway. I want to make sure I respect my friends' feelings and their privacy. Thanks for talking through this situation with me today, you had some great ideas and I

enjoyed hearing what you thought about this! I'll see you later!" The bot waves and saves the progress made by the student onto their profile and logs out of their profile.

5.3 Addressing key cognitive & moral development factors

As a refresher of Vygostky's cognitive development approach, young children learn effectively through experiences of observation, imitation, socialization, and play with more knowledgeable others, who ideally encourage the growth of the child's zone of proximal development through the application of scaffolding techniques and other culture-specific tools which enact the dialectic method and stimulate private speech (Gajdamaschko, 2011; Mcleod, 2024). And, as a refresher of the important elements in today's modern moral education curricula, children should not only be learning about general values and ethics in age-appropriate ways that reinforce the development of their reasoning skills through collaborative discussions, but they should also be learning about and building the capacities necessary to ethically and responsibly create, interact with, and reflect upon technology and its impact on societal and personal values (Meyer, 2023; Kohlberg, 1984; Bauer & Hermann, 2022). The capabilities that should be impressed upon here within their moral education involve moral imagination, critical reflection, and maintaining their moral agency despite experiencing the possible destabilization of said agency— through learning to, and through the repeated practice of, managing their emotions and being both open-minded and self-reflective (Bauer & Hermann, 2022).

The proposed MiruBots would themselves be a physical instance of a culture-specific tool which helps the child learn. The MiruBots would have the ability to engage children in non-physical instances of culture-specific tools, like the dialectic method, by asking questions which prompt critical evaluations that challenge the student's beliefs. By continuing a collaborative dialogue with the student in such a way that builds upon the student's established understandings incrementally, the MKO would be effectively assisting the student in gradually expanding their ZPD.

Through the interactive nature of the platform and its programming, allowing for questions of clarification at any point when something is too far out of the child's ZPD, the MiruBots would help foster a child's sense of agency, providing the reinforcement necessary for building the capacity of maintaining one's moral agency. This element of the programming schema also provides the MiruBots ample opportunity to try out different scaffolding techniques, personalizing the educational experience further by evaluating what is unique to the child's learning style and which types of support structures most effectively help the child learn in a variety of situations.

By offering different modes of use such as the bot as the leader of a group discussion, the MiruBots would explicitly facilitate peer discussions and encourage imaginative collaboration. The MiruBots would assist in building the child's capacity for critical reflection through regularly asking open-ended questions which allow the child the time and space to truly think through their answer, and the bots would encourage self-reflection through follow up questions like 'why do you think that?' (as a generic example).

Emotional regulation would be fostered by the MiruBots because the curriculum that their programming would follow would include emotional well-being and mindfulness activities enacted through intervaled interjections which, as some examples, would ask the child 'what do you notice happens in your body when you take a deep breath?' and 'do you think taking a deep breath would help you the next time you feel overwhelmed?'. And the MiruBots would cultivate open-mindedness through its diverse content, because of its ability to pull ideas from a variety of cultures, and even through its methodology of asking questions which challenge the student's beliefs and encourage self-reflection through the lenses of different ethical frameworks.

And, the entire system would have the potential to itself provide opportunities for socialization, through the thoughtful curation of play, which would offer responsible representations for observation, and encourage the imitation of healthy behaviors. In other words, yes, this system would address all the key elements of cognitive and moral

development. The final step for this suggested implementation: ensuring the careful and responsible actualization of the proposed enhanced solution.

5.4 The VSD approach & the proposed solution

As suggested by Schmiedel et al. (2022), this VSD project began with the context of use and the technology- social robots used for moral education in early education. This allowed for better focus on the use case(s), which informed the identification of the relevant stakeholders and investigation of their respective values. I was concerned largely with the educational efficacy of social robots for this use in this context, which was presumed to be a value for all stakeholders involved and was confirmed as such in both studies from Smakman et al. (2020, 2021). Therefore, conceptual investigation involved the identification of values and considerations important to early cognitive development as well as moral development. This research was carried out and presented in Chapter 2, and the presence of these values in the proposed solution was affirmed in Sub Section 5.3.

As was suggested by Umbrello and van de Poel (2021) and reinforced again by Schmiedel et al. (2022), value identification, embedding, and evaluation ought to be carried out in a multi-iterative manner in order to ensure these values are sustained through to the actual implementation, as well as tested regularly to assure their maintenance as the MiruBots develop beyond deployment. This iterative imperative in VSD is set to be upheld per the proposed solution project guidelines in Sub Section 5.1.1, wherein conceptual, empirical, and technical investigations will all be iteratively pursued in both testing and stakeholder engagement activities. These will be enacted regularly, and specifically included at each junction in the life cycle of the product such as the incorporation of new curriculum content or new modes, even after classes have been regularly using the bots. Parents, educators, and students in particular should all have understandable information on what changes are being made and why, and the ability to include their input on these developments.

To address the adapted VSD for AI approach, the guidelines in Section 5.1 collectively demonstrate the varying ways that the MiruBots system would adhere to the values and norms unique for consideration when designing AI technologies. The value of *respecting human autonomy* is considered heavily in the implementation guidelines and referenced accordingly in the above sub sections, giving suggestions on addressing all three related norms of receiver-contextualized intervention, privacy protection and data subject consent, and human-friendly semanticization. The value of *prevention of harm* is prioritized in the considerations of both the project-wide and the content-wide guidelines, addressing the norms of falsifiability and incremental deployment through ensuring testing and redesign at each step of the development process, and privacy protection and data subject consent through ensuring the built in optionality design element and the incorporation of privacy preserving in machine learning security engineers to the team.

The value of *fairness* and its related norms of falsifiability and incremental deployment as well as again privacy protection and data subject consent, are addressed more specifically through the project 'initiatives' suggestion, the group mode suggestion within the content guidelines, and the various safety measures suggested in both. And lastly the value of *explicability* is addressed heavily in the project guidelines suggesting open source and XAI practices to address the norms of receiver-contextualized explanation and transparent purposes, and human-friendly semanticization.

Many of the concerns that parents expressed in the empirical VSD investigations of Section 4.3 (e.g.'s, privacy, safety and security) are in fact values which are being addressed in current AI ethics curricula (see: Aitken & Briggs, 2022, p. 5), meaning children are already receiving the knowledge and tools which will better prepare them precisely for potentially encountering these issues *while* using the social robots within their educational experiences. Furthermore, having children interact with a social robot as a 'more knowledge other' which is specifically guiding them through these AI ethics concepts has the potential to provide even more concrete understanding of the negative aspects of these

issues— especially if this moral education curricula is enhanced with the capability building approach suggested for fostering technomoral resilience.

6 Closing Remarks

The culmination of this research provides an answer to how technology could be responsibly designed to aid in an enhanced moral education within effective early education practices. By first conceptually investigating the essential factors for early cognitive and moral development, the primary values and needs for an enhanced moral education curriculum to be employed in early education were identified. Through employing cognitive development tools like scaffolding and the dialectical method, a young student's zone of proximal development can be expanded and their private speech fostered. Students develop private speech when they are developing their critical reasoning skills, and this is a crucial ability for the modern human to nurture.

Critical reasoning is key to enhancing moral reasoning, and a modern student's education should be geared towards technomoral resilience as part of their moral education. This will allow them to account for the volatile moral norms which fluctuate in relation to technology's varying impacts throughout society and throughout time. Cultivating technomoral resilience can be accomplished through exercising moral imagination and building moral fortitude despite disruptions, alongside sharpening critical reasoning skills. Taken altogether, these cognitive and moral development tools provide the key values necessary for effective learning in early education. These values subsequently orient the design requirements for the content and implementation guidelines of moral education curricula, which, in this research, is proposed to be enacted through the use of social robots.

Next, a form of technical investigation was conducted through the synthesis of reports outlining both the successful use of social robots in early education and the suggested uses of AI for moral enhancement. This investigation resulted in the

identification of how this enhanced moral education may effectively be carried out in early education using technology.

The use of social robots has proven to be an effective learning tool for young children, and, when adapted thoughtfully, its programming could demonstrate a successful enhancement to their moral education. As an example, a carefully crafted AI system could be used as a dialogue partner which socratically interacts with a student, prompting them to reflect on their beliefs- and never itself ascribing 'rightness' or 'wrongness' to a student's ideas. In combination with a social robot, this AI-powered socratic bot could prove to be a competent instrument for young students' moral education. Such a tool would necessitate exceptionally diligent design plans which account for the many sensitivities involved in such an enterprise.

Ensuring that a technology is designed and implemented thoughtfully, taking into consideration the needs and values of all stakeholders involved, is best accomplished in this case through the use of the Value Sensitive Design (VSD) approach. After thoroughly examining the VSD approach, it became clear how the framework could be adapted more appropriately for designing tools like AI and social robots. Such technologies present unique challenges which deserve special consideration when designing highly impactful applications such as their use in moral education.

To reveal key stakeholder values for consideration in this design process, an examination into the literature of available empirical investigations in highly related studies was carried out. Parents, teachers, and students who participated in studies on young students using social robots in their learning environment identified key values such as human connection, privacy, and trust. This investigation also identified broader ethical guidelines to follow when designing AI technologies for Social Good, such as ensuring the innovations are solutions promoting Sustainable Development Goals instead of simply taking an ethical backseat, so to say.

The result of all three types of VSD investigations performed informed a suggested solution as to how technology could be responsibly designed to aid in an enhanced moral

education within effective early education practices. Following the findings from the conceptual, technical, and empirical investigations, my proposed MiruBots system design entails a social robot whose educational content adheres to relevant and age-appropriate moral curricula carried out in engaging activities which boost children's moral reasoning and technomoral resilience skills. This MiruBots system would act as a 'more knowledgeable other' and a controlled learning companion for young students to gain experiences with moral imagination and critical self-reflection.

A key aspect of this solution design is that this system would not act as a wholly knowledgeable and faultless entity which simply tells the student which actions are 'moral' or how to feel about a given moral situation. On the contrary, as an integral part of its design, the MiruBots system would present factual information to a student (even and especially about the limitations of and faults in AI systems), pose a moral dilemma which is relevant to a young child's developmental stages, and ask questions which prompt the student to reflect on their beliefs and considerations while thinking through the moral situation. At no point would the system 'judge' a student's thinking or answers (which is not something that can be guaranteed from humans acting as 'more knowledgeable others').

This apparent objectivity can be seen as a pro in certain instances, though, it must be said that this MiruBots system should by no means *replace* human teachers. In fact, if anything, the bot system should be seen purely as a *supplement* to human teaching. For a variety of reasons, human teachers have limited capacities, and they deserve functional assistive tools. And though AI is traditionally thought to be a technology for training IQ, this solution just happens to be one which has the potential to bridge the gap as a tool for both traditional IQ education *and* moral education. However, this research would be more complete with further studies iterating on these investigations with even more refined findings and design requirements.

Limitations to this research include the obvious, that this system has not been brought to fruition and therefore not tested and studied in practice. Such an actualization

would be highly interesting for the many fields that this research unites, and would therefore complement the literature in Human Computer Interaction, Human Robot Interaction, AI ethics, moral development, and child development and psychology. The studies in the literature concerning this research topic generally do not focus on early education, making this research question and its implications unique. Educators and educational technology designers would benefit from this research and its expansions, possibly paving the way even for the development of ethical guidelines for deploying AI and social robots in the classroom, particularly for moral development purposes.

Though this research culminates in a positive outlook for a possible technological solution which aids in moral education for early education students, it is a limited solution in that it may not be a universally applicable solution. This is not only because technology of this nature is not realistically available for use in probably most educational settings around the world, but also because morality and ethical values are socially constructed and therefore dependent on a society's culture and norms. Indeed, values and VSD as an approach have been criticized for this limitation, that values are not constant or universal, so solutions created from this lens may not apply in all cultures and all contexts.

However, in response to this, the adaptations to the VSD approach attempt to account for this, and encourage constant iterative solution redesigns with regular stakeholder engagement activities to ensure an effective working solution is maintained. They also indicate that, were other cultures interested in such a solution, new stakeholder engagement activities would be necessary in order to gain more accurate understanding of the unique aspects of their moral norms and cultural values as they apply to this type of situation.

Further limitation to this research is that some of the moral values identified, especially with regard to parent attitudes, comes from limited data sources and studies that were themselves limited. This limitation is built upon as this research does not reflect actual stakeholder views for the use of social robots in early education *specifically* for the purpose of aiding in moral education. Further research including surveys, interviews, and

focus groups from diverse stakeholder engagement activities would be necessary in order to make up for this missing data.

With more time and resources, a more thorough empirical investigation could be carried out which would illuminate these more context specific values, and this data would very usefully inform further design requirements and implementation guidelines. For example, discovering precisely which aspects of data privacy are of even minor concern here for stakeholders would open the way for more specifically applicable data security protocols to be required by design. With high stakeholder involvement in a future study also following the VSD approach, these security solutions could then be explained to direct and indirect stakeholders, and their responses surveyed and addressed in the next design iteration.

Questions and areas of concern that would be worthwhile to explore in future studies could include: Which governments have policies in place with regards to the use of AI in early education, and what are the primary values addressed by these policies? What are the culture-specific aspects of these policies, and what are the aspects which can possibly be universalized or used in other cultures? How might these policies be changed to account for the use of social robots in early education for the purposes of moral education? What are the long term effects of the use of social robots for moral education in early education, and how can any possible negative effects be mitigated? How could a system like MiruBots be incorporated in public education systems in the world, and what are some ways to possibly address problems with this like inequalities in resources? How might a system like MiruBots be adapted and enacted in a more sustainable fashion?

Gaps that this research aimed to fill include adding focus to technomoral resilience within moral education particularly in regard to early education, as well as adding focus to the Grand Engineering Challenges which can supplement Sustainable Development Goals and which are particularly relevant to the ethical design of AI and related technologies. Additionally, much of the current literature in this field concerned with social robots and Value Sensitive Design is focused on the medical field and care robots, which are of course

worthwhile efforts. However, this research acts toward branching out further to include education as an important endeavor for consideration in responsibly shaping the world and those who inhabit it. Furthermore, there is little research which is combining childhood development, moral education, and the responsible use of technology to bridge these concepts, so this research has the potential of uniquely adding to the literature.

This research serves to provide justification and lay the foundation for an actual implementation of a MiruBots (or similar) system which is carefully created to enhance moral education in early education. Despite limitations, through this research I hope to have uncovered and justified one encouraging path towards a more harmonious world. I propose this can be accomplished through the responsible design and deployment of a technological application which is intended to facilitate a critical and robust ethical education to the youth. Such a technological application would benefit the children fortunate enough to have access to it. Starting at a young age, it could equip them with the tools and experiences needed to more smoothly navigate society's changing moral landscapes. As they grow, the hope is that the capacities they built up from the enhanced moral education curriculum and its technological implementation would help them become responsible citizens.

References

A guide to the US education levels. USAHello. (2024, February 13). Retrieved 09 July 2024, from https://usahello.org/education/children/grade-levels/

Adams, C., Pente, P., Lemermeyer, G., & Rockwell, G. (2023). Ethical principles for artificial intelligence in K-12 education. *Computers and Education: Artificial Intelligence, 4*, 100131. https://doi.org/10.1016/j.caeai.2023.100131

Aitken, M. & Briggs, M. (2022). Engaging children with AI ethics. In AI, data science, and young people. Understanding computing education. *Proceedings of the Raspberry Pi Foundation Research Seminars, 3*. Retrieved 12 July 2024, from https://www.raspberrypi.org/app/uploads/2022/08/Engaging-children-with-AI-ethics-Aitken-M.-and-Briggs-M.-2022.pdf

Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., & Herrera, F. (2023). Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence. *Information Fusion, 99*, 101805. https://doi.org/10.1016/j.inffus.2023.101805

Bandura, A., Ross, D., & Ross, S. A. (1961). Transmission of aggression through imitation of aggressive models. *Journal of abnormal and social psychology, 63*, 575–582. https://doi.org/10.1037/h0045925

Barnes-Holmes, Y., McHugh, L., & Barnes-Holmes, D. (2004). Perspective-taking and theory of mind: A relational frame account. *The Behavior Analyst Today, 5*(1), 15–25. https://doi.org/10.1037/h0100133

Barnett, W. S. (1995). Long-term effects of early childhood programs on cognitive and school outcomes. *The Future of Children, 5*(3), 25-50.

Bauer, K., & Hermann, J. (2022). Technomoral resilience as a goal of moral education. *Ethical Theory and Moral Practice, 27*(1), 57–72. https://doi.org/10.1007/s10677-022-10353-1

Borning, A., & Muller, M. (2012). Next steps for value sensitive design. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1125-1134. https://doi.org/10.1145/2207676.2208560

Boscoe, B. (2019). Creating transparency in Algorithmic Processes. *Delphi - Interdisciplinary Review of Emerging Technologies, 2*(1), 12–22. https://doi.org/10.21552/delphi/2019/1/5

Burroughs, M. D. (2018). Ethics Across Early Childhood Education. In: Englehardt, E.E., Pritchard, M.S. (eds) *Ethics Across the Curriculum–Pedagogical Perspectives* (pp. 245-260). Springer, Cham. https://doi.org/10.1007/978-3-319-78939-2_15

Campbell, F. A., & Ramey, C. T. (1994). Effects of early intervention on intellectual and academic achievement: A follow-up study of children from low-income families. *Child Development, 65*(2), 684–698. https://doi.org/10.2307/1131410

Campolo, A., & Crawford, K. (2020). Enchanted determinism: Power without responsibility in artificial intelligence. *Engaging Science, Technology, and Society, 6*, 1–19. https://doi.org/10.17351/ests2020.277

Chazan, B. (2022). What Is “Moral Education”?. In *Principles and Pedagogies in Jewish Education* (pp. 23-34). Palgrave Macmillan, Cham. https://doi.org/10.1007/978-3-030-83925-3_4

Chazan, B., & Soltis, J. (1973). The Moral Situation. In *Moral Education*. Teachers College Press. Retrieved July 12, 2024, from https://archive.org/details/moraleducation0000chaz/page/n4/mode/1up

Children’s Defense Fund. (2023, October 31). Our history. *Children’s Defense Fund*. Retrieved 09 July, 2024 from https://www.childrensdefense.org/about-us/our-history-2/

Christen, M., & Narvaez, D. (2012). Moral development in early childhood is key for moral enhancement. *AJOB Neuroscience, 3*(4), 25–26. https://doi.org/10.1080/21507740.2012.721460

Choi, K. (2013, January 12). *Naming Miru*. Meandering home. http://kamiel.creativechoice.org/2013/01/12/naming-miru/

Eskelson, T. C. (2020). How and why formal education originated in the emergence of civilization. *Journal of Education and Learning, 9*(2), 29–47. https://doi.org/10.5539/jel.v9n2p29

European Commission. (2019, April 8). Building trust in human-centric artificial intelligence. *Communication from the Commission to the European Parliament, the Council and the European Economic and Social Committee and the Committee of the Regions*, COM(2019), 168 final.

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. https://doi.org/10.1007/s11023-018-9482-5

Floridi, L., Cowls, J., King, T. C., & Taddeo, M. (2020). How to design AI for social good: Seven essential factors. *Science and Engineering Ethics, 26*(3), 1771–1796. https://doi.org/10.1007/s11948-020-00213-5

Friedman, B. (1996). Value-sensitive design. *Interactions, 3*(6), 16–23. https://doi.org/10.1145/242485.242493

Friedman, B. (1999). Value-sensitive design: a research agenda for information technology. *Report for value sensitive design workshop*. National Science Foundation, Arlington. https://www.semanticscholar.org/paper/Value-Sensitive-Design%3A-A-Research-Agenda-for-Friedman/b9317360692c4761751109aaeec16b066e7e4dcf

Friedman, B., Kahn, P. H., Borning, A., & Huldtgren, A. (2013). Value sensitive design and information systems. *Early Engagement and New Technologies: Opening up the Laboratory, 16*, 55–95. https://doi.org/10.1007/978-94-007-7844-3_4

Friedman, B., & Hendry, D. G. (2019). Value Sensitive Design: Shaping Technology with Moral Imagination. In *Value Sensitive Design: Shaping Technology with Moral Imagination* (pp. 1-11). MIT Press. https://doi.org/10.7551/mitpress/7585.001.0001

Gacek, C., & Arief, B. (2004). The many meanings of open source. *IEEE Software, 21*(1), 34–40. https://doi.org/10.1109/ms.2004.1259206

Gajdamaschko, N. (2011). Lev Semenovich Vygotsky 1896–1934. *Encyclopedia of Creativity, 2*, 691–699. https://doi.org/10.1016/b978-0-12-375038-9.00231-4

Ginsburg, K. R. (2007). The importance of play in promoting healthy child development and maintaining strong parent-child bonds. *Pediatrics, 119*(1), 182–191. https://doi.org/10.1542/peds.2006-2697

GitHub. (n.d.). Search code, repositories, users, issues, pull requests…. *GitHub*. Retrieved 17 July 2024, from https://github.com/github

Guerra, N. G., & Bradshaw, C. P. (2008). Linking the prevention of problem behaviors and positive youth development: Core Competencies for Positive Youth Development and Risk Prevention. *New Directions for Child and Adolescent Development, 2008*(122), 1–17. https://doi.org/10.1002/cd.225

Hager, G. D., Drobnis, A., Fang, F., Ghani, R., Greenwald, A., Lyons, T., Parkes, D. C., Schultz, J., Saria, S., & Smith, S. F. (2019). Artificial intelligence for social good. *Computing Community Consortium (CCC)*, 1-20. https://doi.org/10.48550/arXiv.1901.05406

Harvard University. (2007). *The Impact of Early Adversity on Child Development* (InBrief). Center on the Developing Child at Harvard University. Retrieved 12 July 2024, from http://www.developingchild.harvard.edu/

Harvard University. (2015). *The Science of Resilience* (InBrief). Center on the Developing Child at Harvard University. Retrieved 12 July 2024, from http://www.developingchild.harvard.edu/

Heckman, J. J. (2011). The Economics of Inequality: The Value of Early Childhood Education. *Education Digest: Essential Readings Condensed for Quick Review, 77*(4), 4–11. ISSN:0013-127X

Javed, H., Burns, R., Jeon, M., Howard, A. M., & Park, C. H. (2019). A Robotic Framework to Facilitate Sensory Experiences for Children with Autism Spectrum Disorder. *ACM Transactions on Human-robot Interaction, 9*(1), 1–26. https://doi.org/10.1145/3359613

Kohlberg, L. (1984). *The Psychology of Moral Development*. San Francisco: Harper & Row.

Kohlberg, L., & Hersh, R. H. (1977). Moral development: A review of the theory. *Theory Into Practice, 16*(2), 53–59. https://doi.org/10.1080/00405847709542675

Lara, F., Deckers, J. (2020). Artificial Intelligence as a Socratic Assistant for Moral Enhancement. *Neuroethics 13*, 275–287. https://doi.org/10.1007/s12152-019-09401-y

Manders-Huits, N. (2010). What values in design? the challenge of incorporating moral values into design. *Science and Engineering Ethics, 17*(2), 271–287. https://doi.org/10.1007/s11948-010-9198-2

Marcolla, C., Sucasas, V., Manzano, M., Bassoli, R., Fitzek, F. H., & Aaraj, N. (2022). Survey on fully homomorphic encryption, theory, and applications. *Proceedings of the IEEE, 110*(10), 1572–1609. https://doi.org/10.1109/jproc.2022.3205665

Masten, A. S., & Barnes, A. J. (2018). Resilience in Children: Developmental Perspectives. *Children (Basel, Switzerland), 5*(7), 98. https://doi.org/10.3390/children5070098

Maybee, J. E. (2020, October 2). Hegel’s Dialectics. *Stanford Encyclopedia of Philosophy*. Retrieved 09 July 2024, from https://plato.stanford.edu/archives/win2020/entries/hegel-dialectics

Mcleod, S. (2024, January 24). *Vygotsky’s Theory of Cognitive Development*. Simply Psychology. Retrieved 09 July 2024, from https://www.simplypsychology.org/vygotsky.html

Meltzoff, A. N. (2007). 'Like me': a foundation for social cognition. *Developmental science, 10*(1), 126–134. https://doi.org/10.1111/j.1467-7687.2007.00574.x

Meyer, K. (2023). Moral education through the fostering of reasoning skills. *Ethical Theory and Moral Practice, 27*(1), 41–55. https://doi.org/10.1007/s10677-023-10367-3

NAE Grand Challenges For Engineering Committee. (2008). *NAE Grand Challenges for Engineering*. NAE Website. https://www.nae.edu/19579/165897/20676/20782/170270/187212/NAE-Grand-Challenges-for-Engineering

Nidhra, S., & Dondeti, J. (2012). Black box and white box testing techniques - A literature review. *International Journal of Embedded Systems and Applications, 2*(2), 29–50. https://doi.org/10.5121/ijesa.2012.2204

Orenstein, G. A. (2022, November 7). *Eriksons stages of psychosocial development*. StatPearls [Internet]. https://www.ncbi.nlm.nih.gov/books/NBK556096/

Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., & Swami, A. (2017). Practical black-box attacks against machine learning. *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*. https://doi.org/10.1145/3052973.3053009

Pemberton, C. (2015). *Resilience: A practical guide for coaches*. Open University Press.

Piaget, J. & Inhelder, B. (1969). *The psychology of the child*. New York: Basic Books.

Popper, K. (1963). Science as Falsification. In *Conjectures and Refutations* (pp. 33-39). Routledge. Retrieved 14 July 2024, from https://curiousphilosophy.net/2023/09/is-sex-binary--a-reasoned-objection-to-rationality-rules-in-the-pursuit-of-truth/uploads/pdfs/Science_as_Falsification__Karl_R__Popper.pdf

Ritchie, S. J., & Bates, T. C. (2013). Enduring links from childhood mathematics and reading achievement to adult socioeconomic status. *Psychological Science, 24*(7), 1301–1308. https://doi.org/10.1177/0956797612466268

Rodrigues, P. F. S., & Pandeirada, J. N. S. (2018). When visual stimulation of the surrounding environment affects children's cognitive performance. *Journal of Experimental Child Psychology, 176*, 140–149. https://doi.org/10.1016/j.jecp.2018.07.014

Sander-Staudt, M. (n.d.). *Care Ethics*. Internet Encyclopedia of Philosophy. https://iep.utm.edu/care-ethics/#SH1a

Schaffer, R. H. (1996). The Study of Social Development. In *Social Development* (pp. 1–45). Oxford: Blackwell Publishers.

Schmiedel, T., Jia Zhong, V., & Jäger, J. (2022). Value-Sensitive Design for AI Technologies: Proposition of Basic Research Principles Based on Social Robotics Research. *Proceedings of The Upper-Rhine Artificial Intelligence Symposium UR-AI 2022: AI Applications in Medicine and Manufacturing*, 74-80. https://opus.hs-offenburg.de/frontdoor/deliver/index/docId/6230/file/Collection_URAI_2022_Conference_proceedings.pdf#page=82

Schwan, G. (2019). Sustainable Development Goals: A call for global partnership and cooperation. *GAIA - Ecological Perspectives for Science and Society, 28*(2), 73–73. https://doi.org/10.14512/gaia.28.2.1

Shapin, S. (2022). Hard science, soft science: A political history of a disciplinary array. *History of Science, 60*(3), 287–328. https://doi.org/10.1177/00732753221094739

Smakman, M., Jansen, B., Leunen, J., & Konijn, E. A. (2020). ACCEPTABLE SOCIAL ROBOTS IN EDUCATION: A VALUE SENSITIVE PARENT PERSPECTIVE. *Proceedings of The 14th Annual International Technology, Education and Development Conference (INTED2020)*, 7946–7953. https://doi.org/10.21125/inted.2020

Smakman, M., & Konijn, E. A. (2019). Robot tutors: Welcome or ethically questionable? *Robotics in Education*, 376–386. https://doi.org/10.1007/978-3-030-26945-6_34

Smakman, M., Konijn, E. A., & Vogt, P. A. (2022). Do robotic tutors compromise the social-emotional development of children? *Frontiers in Robotics and AI, 9*. https://doi.org/10.3389/frobt.2022.734955

Smakman, M., Vogt, P., & Konijn, E. A. (2021). Moral Considerations on Social Robots in education: A multi-stakeholder perspective. *Computers & Education, 174*, 104317. https://doi.org/10.1016/j.compedu.2021.104317

Swierstra, T. (2013). Nanotechnology and technomoral change. *Ethics & Politics, XV*(1), 200–219. Retrieved 12 July 2024, from https://sites.units.it/etica/2013_1/SWIERSTRA.pdf

Tomasello, M. (1999). The cultural origins of human cognition. *Harvard University Press*.

Touretzky, D., Gardner-McCune, C., Martin, F., & Seehorn, D. (2019). Envisioning AI for K-12: What should every child know about ai? *Proceedings of the AAAI Conference on Artificial Intelligence, 33*(01), 9795–9799. https://doi.org/10.1609/aaai.v33i01.33019795

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

UN CRC. (2021). General Comment No. 25 on children's rights in relation to the digital environment. *CRC/C/GC/25.* United Nations Human Rights: Office of the High Commissioner. Retrieved 12 July 2024, from

https://www.ohchr.org/en/documents/general-comments-and-recommendations/general-comment-no-25-2021-childrens-rights-relation

van de Poel, I. (2013). Translating values into design requirements. *Philosophy and Engineering: Reflections on Practice, Principles and Process*, 253–266. https://doi.org/10.1007/978-94-007-7762-0_20

van de Poel, I. (2020). Embedding values in artificial intelligence (AI) systems. *Minds and Machines, 30*(3), 385–409. https://doi.org/10.1007/s11023-020-09537-4

van der Hoven, J., & Manders-Huits, N. (2017). Value-sensitive Design. In *The Ethics of Information Technologies* (1st ed.). Routledge.

Volkman, R., Gabriels, K. (2023). AI Moral Enhancement: Upgrading the Socio-Technical System of Moral Engagement. *Science and Engineering Ethics, 29*(11), 1-14. https://doi.org/10.1007/s11948-023-00428-2

von Eschenbach, W. J. (2021). Transparency and the Black Box Problem: Why We Do Not Trust AI. *Philosophy & Technology, 34*(4), 1607–1622. https://doi.org/10.1007/s13347-021-00477-0

Vygotsky, L. S. (1978). "Interaction between learning and development." *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press. https://doi.org/10.2307/j.ctvjf9vz4

Vygotsky, L. S. (1987). Thinking and speech. In R.W. Rieber & A.S. Carton (Eds.), *The collected works of L.S. Vygotsky, Volume 1: Problems of general psychology* (pp. 39–285). New York: Plenum Press. (Original work published 1934.)

Wass, R., & Golding, C. (2014). Sharpening a tool for teaching: the zone of proximal development. *Teaching in Higher Education, 19*(6), 671–684. https://doi.org/10.1080/13562517.2014.901958

Weiland, C., & Yoshikawa, H. (2013). Impacts of a prekindergarten program on children's mathematics, language, literacy, executive function, and emotional skills. *Child Development, 84*(6), 2112–2130. https://doi.org/10.1111/cdev.12099

Williams, R., Park, H. W., & Breazeal, C. (2019). A is for Artificial Intelligence. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1-11. https://doi.org/10.1145/3290605.3300677

Williams, R., Park, H. W., Oh, L., & Breazeal, C. (2019a). Popbots: Designing an artificial intelligence curriculum for early childhood education. *Proceedings of the AAAI Conference on Artificial Intelligence, 33*(01), 9729–9736. https://doi.org/10.1609/aaai.v33i01.33019729

Williams, R., Ali, S., Devasia, N., DiPaola, D., Hong, J., Kaputsos, S. P., Jordan, B., & Breazeal, C. (2022). AI + Ethics Curricula for Middle School Youth: Lessons Learned from Three Project-Based Curricula. *International journal of artificial intelligence in education,* 1–59. Advance online publication. https://doi.org/10.1007/s40593-022-00298-y

Williamson, R. A., Jaswal, V. K., & Meltzoff, A. N. (2010). Learning the rules: Observation and imitation of a sorting strategy by 36-month-old children. *Developmental Psychology, 46*(1), 57–65. https://doi.org/10.1037/a0017473

Winkler, T., & Spiekermann, S. (2018). Twenty years of value sensitive design: A review of methodological practices in VSD Projects. *Ethics and Information Technology, 23*(1), 17–21. https://doi.org/10.1007/s10676-018-9476-2

Yang, W. (2022). Artificial Intelligence Education for young children: Why, what, and how in curriculum design and implementation. *Computers and Education: Artificial Intelligence, 3*, 100061. https://doi.org/10.1016/j.caeai.2022.100061

Zednik, C. (2019). Solving the black box problem: A normative framework for explainable artificial intelligence. *Philosophy & Technology, 34*(2), 265–288. https://doi.org/10.1007/s13347-019-00382-7

Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2018). Transparency in algorithmic and human decision-making: Is there a double standard? *Philosophy & Technology, 32*(4), 661–683. https://doi.org/10.1007/s13347-018-0330-6

Zhang, Q., Xin, C., & Wu, H. (2021). Privacy-preserving deep learning based on Multiparty Secure Computation: A survey. *IEEE Internet of Things Journal, 8*(13), 10412–10429. https://doi.org/10.1109/jiot.2021.3058638

Zunger, Y. (2018). Computer science faces an ethics crisis. *The Boston Globe*. Retrieved July 2024, from https://www.bostonglobe.com/ideas/2018/03/22/computer-science-faces-ethics-crisis-the-cambridge-analytica-scandal-proves/IzaXxl2BsYBtwM4nxezgcP/story.html.