Applying Dietrich Bonhoeffer’s Christian Conception of Responsibility to the Ethics of Artificial Intelligence Design and Development

Joseph Francis Nelson

Submitted in accordance with the requirements for the degree of a Masters by Research in Theology

University of Leeds / Leeds Trinity University

School of Children, Young People, and Families and the School of Computer Science

March 2023

The candidate confirms that the work submitted is his own and that appropriate credit has been given where reference has been made to the work of others.

Acknowledgments

I would like to thank my supervisors, Dr Antesar Shabut and Dr Suzanne Owen, for their support, patience, and guidance. I would also like to thank Dr Ann Marie Mealey for her time as my supervisor in the early stages of this project and for always encouraging me on my research and educational journey.

I would like to thank my friends and family and everyone who has helped and supported me in my PGR journey. For my friends Hannah and Adam Knight, and Laura Whitaker. I would especially like to thank my Mother, my sisters Clare and Laura, my Grandmother, and my Uncle. I would also like to give special thanks to Dr Anne Marie Sowerbutts and Rosie Hodnett.

My Church community and congregation have also been a great support to me during this time especially Bishop Walter Jagucki and Pastor Mark in Leeds along with the rest of the clergy of the Lutheran Church in Great Britain.

There are many other people I could thank. Most of whom I cannot name here. It has been quite a journey over the pandemic to complete this thesis and I am indebted to all those who have supported and pushed me to continue despite difficulties and hardships.

Abstract

This thesis will consider the ways in which Dietrich Bonhoeffer's ethical conception of responsibility can be applied to the current discussions around the ethics of the development and implementation of Artificial Intelligence technology. As we will see, the topic of AI ethics is not a question that any of us can ignore. It pervades modern technology and influences our modern world. Far from seeking to stifle technological development this thesis seeks to promote ways in which we can use these technological developments responsibly and find ways to develop AI for the good of all humanity.

This thesis will begin by considering what Bonhoeffer means by ethical responsibility and how that can be interpreted and implemented in our modern ethical landscape and applied to the world in which we find ourselves today. Furthermore, this thesis will see what sort of governmental frameworks and recommendations already exist and how there are pitfalls that need to be addressed by a wider ethical hermeneutic of responsibility.

We will also consider what AI is and briefly outline some key considerations before applying Bonhoeffer's ethics of responsibility to the concepts discussed.

Finally, this thesis will conclude that Bonhoeffer's ethics of responsibility can help Christians, and those outside the Church, to think of the current debates around Artificial Intelligence in ways that go beyond the current pitfalls of a too heavily principle-based argument or one that is more concerned about cataclysmic questions of AI taking over the world. Bonhoeffer points us back to the here and now, to that which is real, and the ways in which we can, as individuals, work for a more just and fairer society which stands up for the weakest and most vulnerable.

Table of Contents

Table of Figures

Introduction

This thesis will consider the writings of Dietrich Bonhoeffer on the topic of ethical responsibility and how his thoughts can be applied to the ethics of Artificial Intelligence development and implementation. Bonhoeffer's ethical reflection was produced at a time, which like ours, was filled with great change and rapid technological development: the Second World War. His context helps us to understand the risks that we face today from a similarly rapid environment of technological development. For example, we are seeing a change in the way that media works today and how it can be influenced by AI and other forces. Likewise, Bonhoeffer saw, in his day, how new and modern media was used by the Nazi propaganda machine to influence individuals and groups of people. Bonhoeffer is significant in his theology of responsibility because he seeks to 'find a way of talking about responsibility that does not collapse into individualism' (Reed, 2018, p87). As Keith Clements points out; theologians and a wide range of people today are constantly finding more and more in Bonhoeffer, and he speaks to the situations that we face in a way that many other theologians do not (2022, p1-2). Bonhoeffer is particularly relevant for the witness of his life, and for the way his writing 'fuses a Christ-centred view of revelation (owing much to Karl Barth) with an intensely relational and social understanding of both the church and humanity at large' (Clements, 2022, p5). Relationships are key to Bonhoeffer's theology, as is community, both of which are lacking in the modern world. Bonhoeffer is useful in our discussion on AI because it is through a rediscovery of our communality with all persons, and as a result rediscover individual responsibility, that we can ensure that AI is a tool that is used for the betterment of society and humanity.

As anyone who reads the news today will be aware: the recent technological development of Artificial Intelligence is one of the most pressing and relevant ethical topics of the modern era. This technology is already changing how we live and has the potential to do so to greater degrees in the future. For example, AI is used in, mobile phones, medicine, crop rotation, and social media to simply name a very small number of fields where it is already prevalent. AI is already being used in one way or another across the spectrum and variety of life.

It can be argued that Artificial Intelligence is, by its very nature, 'a cross-disciplinary approach to understanding, modelling, and replicating intelligence and cognitive processes'

(Frankish and Ramsey, 2014, p1) and thus requires a range of perspectives and inputs from a variety of sources and disciplines. This technology has the potential to, and indeed already does, affect, and effect, our lives on a daily and even hourly basis. Thus, theology, philosophy, and in particular, the discipline of ethics which these other disciplines feed into, cannot ignore the need to be attentive, responsive, and even involved in the development and implementation of AI.

In this context it is important to note that ‘from the very outset of his theological development, Bonhoeffer asked how systematic theology can be developed today to contribute constructively to [the] academic, societal, and ecclesial discourse’ (Harasta, 2014, p15) of his time. Dietrich Bonhoeffer sought to develop a theology of daily life which was not the domain solely of elites in the academy but that was also accessible to and useful for the ordinary Christian in their journey of discipleship. Bonhoeffer saw the Church as being like a child that is constantly learning and growing and developing before God. He did not see the Church as a great monolith, unchanging, separate from the world in which she finds herself. Rather Bonhoeffer saw the Church as the *Sanctorum Communio* (1930) in which Christians move together in expectation of the fulfilment of the Christian hope in Jesus Christ. ‘Bonhoeffer's interpersonal ethic of dialogic confession reaches out to others while firmly surrounded by a story that connects one to others through ‘bridges’ of service and awareness of other perspectives’ (Arnett, 2005, p219). Therefore, the Church cannot ignore the pressing concerns of life and the world in which she finds herself. The Church has a place and role in society by which it calls for the common good of all persons and for justice: and the same is true in the context of Artificial Intelligence.

Unlike our modern worlds individualistic context and worldview, Bonhoeffer’s ‘interpretational ethic is neither an absolute unleashing of individual liberty nor a desperate clinging to an old moral system, but an interpersonal life guided by a centre that reaches out to others’ (Arnett, 2005, p219). Individualism has no place in Bonhoeffer’s ethical conception but likewise Bonhoeffer stands in opposition to state overreach and emphasises that individual responsibility is of central importance to the life of the Christian. Likewise, it is important to note that for Bonhoeffer individual responsibility is always at the service of the community. Indeed, it is not unreasonable to refer to Bonhoeffer’s (Lutheran informed) conception of ethics as a whole as ‘an ethic of responsibility’ (Nissen, 2011 B, p103).

Responsibility, for Bonhoeffer, recognises that which is real. For example, when it comes to a similar issue to Artificial Intelligence - climate change - that I alone, as an individual, can do very little to impact the environment but as I act responsibly, I can be a voice for change and play some part in the common good by being responsible for what I do and what I consume to the best of my ability. Yet to be responsible is also to recognise that there are things that I cannot control.

Moral Responsibility calls us to go beyond law and principle, not by disregarding these things but by discerning them, we are to internalise the moral law and then to apply it in a way that recognises the reality and complexity of the situation in which we find ourselves. Responsibility responds to the reality of the situation - not the ideal - by taking responsibility for the decisions that we make and by taking responsibility not just for ourselves but also for others.

It is clear that 'Trust in God's omnipotence and act in the world does not contradict the thought of humans' own responsibility. Moreover, faith, as the unconditional trust in God, sets persons free to assume responsibility and to act in freedom. Responsibility acting out of faith does not seek to replace God, but instead dares to trust in God's action within the world. Talking of God today should strengthen people in their faith and trust "that God is not just timeless fate, but waits for and responds to sincere prayer and responsible actions"' (CPCE, 2022, #69).

This thesis seeks to answer the question of how Dietrich Bonhoeffer's concept of responsibility could be related to the ethical questions posed by the design and development of Artificial Intelligence. This is necessary because Artificial Intelligence is a modern ethical issue that defies our normal, previous, western ethical concepts related to Intention- Act – and Agent (Reed, 2018, p87). There is rarely, in such modern issues, a direct causal link between my action and the consequences of those actions. For example, in retail- by buying certain products I may be, in part, the cause of the existence of a sweat shop in a third world country. Yet the consequences of my actions, due to the globalised world that we live in, are not as evident as they would have been in the more traditional socio-economic systems of the past. Likewise, Artificial Intelligence technology has far-reaching consequences that are far from straightforward when it comes to ethical decision making. As with many modern issues - traditional ethics fails us because it is not designed to confront the extraordinary situations in which we find ourselves today. Especially when we consider that the extraordinary has too often become the ordinary. In AI we are no longer creating something that will simply do what we tell it but something that has the potential to go beyond our mental capacities in ways that we have not seen before.

Bonhoeffer too faced a time of rapid technological development and social change and recognised that the ethical frameworks of his time and times past had failed in their attempts. Bonhoeffer watched as the Nazis used radio and cinema as propaganda tools to push their message and to facilitate a destructive and genocidal war upon the world. In response to this Bonhoeffer pointed to the need for us to take up responsibility.

Theology, philosophy, and ethics do have something to say when it comes to the ethical questions surrounding Artificial Intelligence technology. This is particularly true when it comes to theology. Theology is after all ‘a reflection on God and on the revelation given by God to human beings’ (Arroyo, 2002, p68). Yet, just as with humanities, this discipline does have important things to teach us in our technological age. Technology impacts every area of life. But so do questions of theology, philosophy, and ethics. The fact is that these disciplines have shaped how we think and act, especially for Christians; but also, all people alive today have also been shaped in some way by these disciplines. Therefore, it is not unreasonable to suspect that there might be something of use in these disciplines that we can learn from. We live within a cultural and historical context and therefore the humanities are essential in developing Artificial Intelligence for the common good of all persons.

We need an ethics for Artificial Intelligence precisely because of the harm that it could do to privacy, rights, jobs, and many other things. However, even more than this- for a technology to be successful it needs to be something that people can trust. With higher risks to individual privacy, jobs, and freedoms, there needs to be a higher standard of responsibility. Accountability is something that humans, such as developers and creators, may rather avoid; especially when it comes to a technology like this one where there are so many extensive variables.

There have been many attempts by governments and other organisations, as we will see in later chapters, to develop an ethical framework for Artificial Intelligence development and implementation. Many of these are incredibly beneficial and good tools. However, there are two common themes or weaknesses that will also be evaluated later in the thesis, and these are that such frameworks tend to either:

a) Focus too much on future possibilities and risks, or

b) So principle based that they do not actually translate into action.

There has already been some Christian reflection on the topic of Artificial Intelligence including several books and the historic *Rome Call for AI Ethics* (2020), which I discussed in the paper *The Rome Call to Artificial Intelligence Ethics Inside the mind of the Machine: How the Church can respond to the ethical challenges presented by AI* as part of my research conducted during the writing of this thesis and will not discuss in detail here (a copy will be provided with the thesis). However, these books and agreements too often fall into the difficulties given above.

For example, John Lennox's (2020) book, *2084: Artificial Intelligence and the Future of Humanity,* gives significant time to his consideration of Scripture as a tool for the Christian in their assessment of the current and future directions of Artificial Intelligence ethics, development, and deployment. Lennox looks at passages from Genesis to Revelation, within a Christocentric and Eschatological framework culminating in the claim of Jesus as the supernatural superintelligence (2020, p158). The central argument beyond all else for Lennox is that death is not simply a 'technical problem' (2020, p159) that needs to be solved and thus in like manner, the aims of Artificial Intelligence development must be about making us more fully human and never contribute to the reduction of our humanity or dignity. In the end we must recognise that our aim cannot be to make what Lennox calls through his work the 'homo deus' or superintelligence (Lennox, 2020, 227) for there is 'no way to a glorious future that bypasses the problem of human sin' (Lennox, 2020, p227).

Lennox presents some stark warnings when it comes to the potential future of Artificial Intelligence and his assessment of the dangers of an Artificial General Intelligence or Superintelligence should not be taken lightly. However, as he himself says that at present 'AI and machine learning algorithms are … no more alive than Microsoft Word' (Lennox, 2020, p96). Thus, there is also a need to consider a scriptural response to current AI uses and development. The issues around surveillance, privacy, and jobs along with other ethical dilemmas are arguably more pressing. Lennox does discuss 'violations of human rights' (2020, p 72) AI controlled weapons (2020, p74) and the risk of job losses (2020, p64). However, it is also notable that his scriptural exegesis is confined, for the most part, to his analysis of superintelligences and potential future threats. This is particularly concerning when the more pressing issue for our ethical consideration must be the dignity of the human person. We must remember that the threats that are posed by Artificial Intelligence to the human person already exist and are often being ignored in favour of long-term reflection by theologians and philosophers.

Just as with Lennox, Noreen Herzfeld, in her book *In our Image: Artificial Intelligence and the Human Spirit* (2002) begins the scriptural analysis of Artificial Intelligence with a reference to the book of Genesis and contemplating at first humanity as being made in the image and likeness of God (Herzfeld, 2002, pp 10-11). The main reason that so many writers start with this has been noted by a range of authors- the question of whether there is a conflict 'between AI and biblical teaching about the… human soul' (Schuurman, 2019, p5). Humanity's place as the 'imago Dei' (Herzfeld, 2002, p10) is central to many Christians, such as in Catholic Social Teaching, with regard our dignity and the respect due to human life. This leads us to question what Artificial Intelligence is… if we are the image of God and we make an artificial life that is made to be in our image then what is AI in terms of theology? However, this in and of itself is far from simple.

What does this have to do with Artificial Intelligence? It can be argued that in creating an artificial intelligence that imitates us and works in a similar way to us is that we are either striving to make AI in the image of God (imago Dei) or in the image of ourselves (imago Hominis). In many ways AI has the potential to do this far more than any previous technological development because it is coming to conclusions without external human direction. It is in some ways like a human or like God. The idea of making AI in the image of God or in our own image is not a scripturally justifiable aim. To make something to be like a god is idolatry (Ex 20:4). Furthermore, Herzfeld argues that it is our being made in the image of God that makes humanity distinct from the rest of creation in Gen 1 and Gen 9 (2002, p15). Thus, anything made in our image would be nothing more than a cheap representation of the image of God and thus likewise arguably idolatrous. Instead, Artificial Intelligence must be a tool that we use rather than a quest for something like or greater than us. The human heart has a longing for God – or for 'another intelligence' (Herzfeld, 2002, p94). Yet, this longing that many seek to fill with technology is arguably truly a longing for God. Yet our longing for God can never be truly intellectual because it is in relationship that we truly encounter God and not by intelligence alone.

Just as with Lennox, the issue in Herzfeld's writing regarding a scriptural perspective on Artificial Intelligence is that, as she herself points out, 'we do not yet have intelligent computers' (Herzfeld, 2002, p 94) in the way that she is discussing. Indeed, the argument is that our aims, objectives, and limitations must be an essential part of any ethic of Artificial Intelligence. We must not seek to create Artificial Intelligence as anything other than a tool to enrich human life. However, this ignores something that is arguably more central to the Christian Gospel and the Word of God- to care for those in need here and now (Matt 25:35-36). As important as it is to contemplate future developments and potential uses, we need a scriptural ethic that is looking at how Artificial Intelligence is being used here, now, and today. As both books demonstrate- there are considerations around the future of Artificial Intelligence and their impacts and upon transhumanist uses. However, we are lacking a scriptural assessment of current Artificial Intelligence developments, and it is all too tempting to focus simply on the future possibilities rather than the current situation.

Therefore, this thesis will seek to establish what Dietrich Bonhoeffer means by responsibility and what that term implies for developing an ethical approach to Artificial Intelligence design and development (Chapter1: *Bonhoeffer's Conception of Responsibilit*y). Later, in this document, (Chapters 2: *What is Artificial Intelligence and how does it work?* and Chapter 3: *Current Codes of Ethics in use today*) will consider what Artificial Intelligence is, the ethical questions it presents, the current attempts to develop an ethics of AI. In Chapter 4 (*Issues and Applications of Bonhoeffer),* we will consider some of these issues and look at how we could apply Bonhoeffer's ethics of responsibility to the context of Artificial Intelligence technology. At the end of this thesis, we will conclude to what extent this approach has been useful in the ongoing quest to develop an ethics of Artificial Intelligence.

Chapter 1: Bonhoeffer's Conception of Responsibility

1.1 Introduction

Bonhoeffer's ethics of responsibility is all about real life. It goes beyond the categorisation of norms and principles, and other such abstract ideas. Yet he also rejects any notion of individualism, situationism, or consequentialism. Instead, Bonhoeffer's ethics of responsibility is all about encounter with human beings and the reality of the world. This is precisely because, in Bonhoeffer's theology, Christ is the ultimate norm of ethical principles and is the ultimate reality. Therefore, responsible living is always a response to the other, and preference for the other, rather than a self-centredness. In other words, responsible living is to live as Christ lived - for others.

Dietrich Bonhoeffer was a 20th Century Lutheran theologian, pastor, academic, and a member of the resistance against Hitler within Nazi Germany. He is well known as a Christian Martyr and theologian who has had a wide-ranging impact upon 20th and 21st Century Christian theology. He grew up in a home steeped in a 'great historical heritage and intellectual tradition' (Nelson, 1999, p23) and his love for the world of thought and theology continued throughout his life including his doctorate, which was completed before he was 25. His academic career began when he was just 24 (Nelson, 1999, p29).

Christology was central to Bonhoeffer's theology and his subsequent academic and social work. Likewise, Christology forms the backdrop to his conceptions about responsibility in ethics.

Bonhoeffer was a member of 'the anti-Nazi Confessing Church' (Floyd, 2009, pvii). He was certainly a leading light in theology in his lifetime and would have been for much longer if it had not been for his premature death. *Ethics* (1949/ 2009) is one of Bonhoeffer's final works and was unfinished at the point that he was executed. This is the text upon which much of our discussion about Bonhoeffer and responsibility will be based. The fact that the book was unfinished at the point of his death significantly complicates the investigation that this paper seeks to undertake as it is unclear how Bonhoeffer would have integrated his conception of responsibility into a wider ethical framework. Likewise, some of the clarification that could have been added such as the boundaries of responsibility also have not been elaborated as much as could have been desired. For example, he writes down a possible chapter title of *Responsibility as Love* and yet went no further. What he means by this is not clear.

Bonhoeffer is significant in that his understanding of ethics and Christian life is responsive to the situation in which the Church finds herself. Indeed, life and situations have evolved since his time, however, the desire to see a Christianity that is always being attentive to the 'signs of the times' (Gaudium et Spes, 1965, #4) has become a prevalent theme across Christian denominations in recent times. Bonhoeffer recognised that traditional conceptions of Christian ethics had failed the world and led to the Church being passive to the threat of national socialism. His works, therefore, (near the end of his life) tend towards a vision of Christianity that looks to a future that builds a theology that stands for justice and right but that is also responsive, less static, and more focused upon the person of Jesus Christ. The three themes of Bonhoeffer's theology are: 'discipleship and community,' the 'struggle for peace and justice,' and finally 'faith in a secular age' (Gruchy, 1999, p103). All three of these themes are highly relevant for the world today and theology post the Second World War. In many ways one could argue that Bonhoeffer was ahead of his time. While his work does not necessarily corollate directly onto the struggles that we face today (such as Artificial Intelligence) we can still discover 'different trajectories in his legacy which help Christians to engage in the struggle for justice and liberation more faithfully' (Gruchy, 1999, p104). Bonhoeffer challenges Christians and others to faithfully engage with the world and to not be afraid of moral peril or guilt that might result in this but to embrace the fact that things are messy for the sake of love and, for the Christian, for the sake of the Gospel. This is especially relevant today when it comes to Artificial Intelligence as it is a topic that many Christians, and others, are unwilling to engage with.

While Bonhoeffer's ethics might seem, to a non-Christian, to be irrelevant outside the scope of Christian theology and morality - it is important to remember in this chapter that 'Bonhoeffer's work predates the emancipation and expansion of ethics' (Harasta, 2014, p14) and during his period 'the primary basis for ethical claims remained dogmatic theology' (Harasta, 2014, p15). Therefore, the form of interdisciplinary study that we are engaged in today would not have been normal while he was alive. It could be argued therefore that there is a scope for a Bonhoeffer based ethics of responsibility that, while recognising the importance of Christianity and the person of Jesus Christ in his work, goes beyond what Bonhoeffer says and makes his concepts applicable even outside of the disciplines of Dogmatics and Moral Theology. Making Bonhoeffer accessible and relevant even to the non-Christian and therefore relevant to our modern considerations of the ethics of Artificial Intelligence.

When it comes to Artificial Intelligence there is a gap. This gap is not common in traditional ethics. The distinguishing factor is that there is a separation between actor and consequence. This issue is becoming more and more common in our globalised world. The question of *who is responsible* is not clear or recognisable within AI, in the sense that the responsible agent has been in the past, because even though Artificial Intelligence is programmed by a person - we cannot always fully understand or comprehend how or why the machine has reached the conclusions that it has. Therefore, as we explore these new areas of ethics and technology it is essential that we take up the challenge to question what it means to be a responsible agent and how the responsible agent is to act ethically in the modern world – especially, for this study, when it comes to the questions of ethics and Artificial Intelligence.

This chapter will seek to set out what Bonhoeffer understands moral responsibility to mean and how he interprets the call to be a responsible agent. Furthermore, this chapter will go on to look at who that moral agent is and whether, or not, one is required to be a Christian to be a genuine moral agent as Bonhoeffer seems to suggest at some points. To situate Bonhoeffer within his context this chapter will consider the wider Lutheran understanding of ethics and responsibility from which Bonhoeffer emerges and question how this might contribute to the discussion around the responsible agent. It is important to note that Bonhoeffer is not the only thinker to consider the question of moral or ethical responsibility and therefore this essay will also consider other perspectives on the question and how these different perspectives might contribute to our discussion. To conclude this chapter, we will consider how we will frame a Bonhoeffer based conception of ethical responsibility that can be used to inform our considerations of the ethics of Artificial Intelligence.

1.2 Lutheran Context

As with Bonhoeffer, the start point of Lutheran Ethics are different to the normal start points of philosophical and theological ethics. Lutheran ethics, as with all Lutheran theology, starts with the belief that to aim to be good enough to justify one's own salvation is a fool's errand. Lutheran theology begins from a starting place of recognising that we will fail, and no ethics, legal system, or law will ever be sufficient. In the words of Lutheran theologian Steven D Paulson:

'Lutheran theology begins perversely by advocating the destruction of all that is good, right, and beautiful in human life. It attacks the lowest and the highest goals of life, especially morality, no matter how sincere its practitioners' (2011, p1).

This is precisely because the starting point of Lutheran theology is faith rather than works - and the source of that faith which is the revelation of God in Christ. This, as we will see, is very prevalent and a core element of Bonhoeffer's ethics.

Lutheran ethics is not about being good or about becoming a good person. We cannot do these things. Instead, one could argue that Lutheran ethics is about, and what one could argue Bonhoeffer's ethics is about, how are we to live authentically in a broken world from within a Christological context.

Salvation, for Lutherans, has nothing to do with being good but instead has everything to do with recognising that we can never be good and that we need to place our faith in Christ as the only one who can make a way for us to be saved. No matter what we do we can never be good enough.

However, in history, this doctrine has consequently helped to shape the world in which we live today. Bonhoeffer argues that 'Luther's great discovery of the freedom of the Christian and the Catholic heresy of essential goodness of human beings resulted together in deifying humanity' (Bonhoeffer, 1949/2009, p123) and that this led to the enlightenment. He goes on to argue that the cause of the enlightenment, and the secularisation that has slowly come about because of the enlightenment, has resulted in individualism which Bonhoeffer sees as antithesis to the Christian message. This individualism finds its greatest expression in the allergy amongst many to talking about faith or politics; or God forbit the Church making a statement about the state which, even today, can be looked down upon by many as the Church stepping out of line to a place that it has no business commenting on. This protest against individualism is a key factor in Bonhoeffer's ethics of responsibility as we will see later and an important element in developing a non-individualistic ethics for Artificial Intelligence technology.

A central element of Lutheran Theology and Ethics is the concept of *The Two Kingdoms*. This idea arises out of the question of the place of the law/ God's law now that we have removed the moral law from the sphere of salvation. Lutherans recognise that there is a place for the law and that it cannot just be ignored - the teaching of salvation by grace alone through faith alone is merely saying that we cannot sufficiently keep the law. Lutheran theology is not saying that the law isn't important.

In Lutheran theology the word law can have multiple meanings. Firstly, the law (in this case as found in the Old Testament or Hebrew Bible and the law of God written in our hearts and consciences) shows us our sin and to lead us to repentance. Secondly, the law can be used in our attempt as believers for our growth in holiness. Lastly, there is the law of the civil realm. These three forms of law are also not identical to one another: indeed, they can contradict one another.

While the content of the laws of faith, of both the Old and New Testament, may be different to the laws of the state, Lutheran ethics does point to the importance of law when it comes to state law as being significant and important for the Christian. This usage of the law is important as it demonstrated, during the reformation era, the obedience to the state of these rogue protesting religious rebels – however, it did further entrench the view of the two kingdoms theology.

As a result, the Lutheran Augsburg Confession states that 'all government in the world and all established rule and laws were mandated by God for the sake of good order' (AC Article XVI) [1]. Likewise, the Augsburg Confession teaches that 'Christians are obliged to be subject to civil authority and obey its commands and laws' (AC Article XVI) except when obedience to the state would be sinful. This is based on the concept found in Romans 13:1 where Paul is talking about the importance of obedience to the ruling authorities. In this concept of the two kingdoms- where in one 'kingdom' we have God where we enter the promise by faith; the

[1] Augsburg Confession

second is where ‘God uses the law to establish external peace’ (Paulson, 2011, p251) using civil authority.

As such we see early on in Lutheran theology the start of the distinct separation of Church and State. In this we see an issue beginning to develop. Luther recognises both the Church and the State as being established by God with two different mandates - the Church to preach and teach the Gospel - the way to salvation - and the state to maintain order in society. The Church developed a distinct allergy to discussing the state and what the state should do precisely because the mandate of the Church is different to the mandate of the state. However, this was to have disastrous consequences. We see this very clearly in Bonhoeffer’s time during Nazi rule in Germany. There was an unwillingness on the part of many Christians to engage in resistance against the state as they regarded this as unchristian. Whereas Bonhoeffer argues that to do nothing is the thing that is unchristian. Instead, Bonhoeffer, while maintaining an awareness of the different mandates of Church and State, also sees the role of the Christian as being to make the Kingdom of God present on earth. Instead of the Church/State relationship being about two fully separate entities - the role of the Church is to recognise the call of God in the situations in which they find themselves and to respond - whether that means acting with the state or against the state; and similarly, the Church and Christians have an obligation to speak about the issues facing us today- in this instance Artificial Intelligence.

In the book, *Free in Deed: The heart of Lutheran Ethics,* Craig Nessan argues that ‘history has been tragically marred by the failure of the Church of Jesus Christ to actively oppose political tyranny’ (Nessan, 2022, p57). Nessen speaks of the Two Kingdom doctrine and the ways in which it has fallen short in calling the Church to stand up to tyranny and evil. To overcome the issues inherent to the two-kingdom doctrine Nessan rephrases Luther's teaching on the two kingdoms in a similar way to Bonhoeffer. When considering the two Kingdoms we are to remember that ‘finally there is only one kingdom of God’ and that this kingdom ‘broke into the world’ (Nessan, 2022, p59) through the incarnation of God in Jesus Christ. Like Bonhoeffer, modern Lutheran ethics focuses more on the kingdom coming on earth as it is in heaven (Matthew 6:10); as opposed to an ethics of a distinctiveness of Church and state. With Bonhoffer, modern Lutheran ethics, in the concept of the two kingdoms, is no longer saying that the Church should not comment on the state but rather that the political world is to be called to responsibility by the Church, and that the call to responsibility is ‘one of the most urgent theological tasks belonging to Lutheran ethics’ (Nessan, 2022, p65). In the end the role of the Church is to bring forth ‘the one kingdom of God’ (Nessan, 2022, p65) and that Christians ‘serve God by living responsibly’ (Nessan, 2022, p65).

In conclusion, Bonhoeffer is formed by his Lutheran context and as such places Christ and justification by faith alone through grace alone at the heart of his conception of what it means to be a Christian and thus at the heart of his ethics. Bonhoeffer rejects the notions found during his lifetime relating to the two kingdoms and Church and State and has influenced Lutheran ethics going forward. Bonhoeffer, and subsequently modern Lutheran ethics, instead calls for an ethics of responsibility and accountability to pervade ethical decisions and politics. This is significant in our discussion about Artificial Intelligence because it means that the Church and Christians in general have an obligation to be informed and involved in developing the ethics of AI and not to ignore it.

1.3 Bonhoeffer’s Conception of Responsibility

In the early 20th Century, there were a wide range of differing ethical conceptions. However, the concept and theme of ethical responsibility was increasingly emphasised and discussed amongst ethicists and theologians - particularly in the context of war and the blurring of moral boundaries of the past (Lovin, 2016, p66). Increasingly, there was a recognition that traditional Christian morality was no longer working or able to engage with the reality of the situation in which people now found themselves. As a result, Bonhoeffer began to write his book *Ethics* (1949/2009) and to explore what it meant to be a Christian living and acting in the world in his time.

The centre of Bonhoeffer’s theology and ethics is always the person of Jesus Christ. This is where his book *Ethics* (1949/2009) starts and is the central theme throughout. He argues that ‘Responsibility, in the Biblical sense, is primarily a response, given at the risk of one’s own life, to the questions people ask about the Christ event’ (Bonhoeffer, 1949/2009, p255). Bonhoeffer believed that ‘taking responsibility occurs before God and for God, before human beings and for human beings; it is always answering and being responsible for the sake of Jesus Christ’ (Bonhoeffer, 1949/2009, p 256). In other words, we are responsible, and our responsibility is firstly a response to Jesus but also, we are responsible not just to our neighbour but before God and before humanity.

Likewise, Bonhoeffer argues that ‘responsibility is based on vicarious representative action… This is most evident in those relationships in which a person is literally required to act on behalf of others’ (Bonhoeffer, 1949/2009, p 257). Vicarious responsible action is entirely about taking the place of the other or acting in the place or on behalf of others. It echoes the concept of being responsible for the sake of Jesus because it means doing what Jesus did - taking on the responsibility for others and bearing the guilt of others even when we are innocent. In his conception of responsibility, particularly in the case of vicarious responsible action, Bonhoeffer is trying to move his readers conception of ethical responsibility from a legalistic framework to a form of ethical reflection that moves his reader ‘beyond individualism’ (Arnett, 2005, p85), and indeed the individual, and encourages them to instead consider ethics as an action that flows from an expression of relationship.

As we can see from the above quotes Bonhoeffer is not necessarily espousing what would normally be considered a traditional law based (or deontological) Christian ethics. While many theologians, such as Gerard Hughes, do agree with Bonhoeffer that ‘the ultimate norm for Christian belief is God’s revelation of himself in Jesus of Nazareth’ (1991, p7) and that this therefore also applies to our conception of ethics: other theologians such as Thomas Aquinas would argue that, when it comes to ethics, there is a Natural Law, that is discernible to the individual apart from divine revelation, that should be the basis of our ethical decision making (1485/1991, p107). It is not necessarily the case that these two ideas /ethical systems are necessarily in conflict with one another. However, it could be argued that at times Bonhoeffer’s concept of responsibility, could require Bonhoeffer’s reader to take actions that may (potentially) be seen as being in breach of natural law. For example, this could be the case where Bonhoeffer was involved in by attempting to assassinate Adolf Hitler and thus taking on responsibility and guilt. For example, in deontology/Natural Moral Law murder or the assassination of a political leader would always be wrong- Bonhoeffer does not deny this- but what he does say is that in Christ we can act responsibly and take on guilt through vicarious responsible action[2].

This is strongly related to the question of the source of Christian morality which is an ongoing debate within Christian Churches. Some argue, as with Natural law, that ethics is something that is universal or in other words - it does not matter who you are the moral law is the same and is able to be reached through reason, separate from revelation. Others argue that there is something specific and particular about Christian ethics that is either not accessible to those from outside the Church or not applicable to those outside of it. Moreover, others would argue that ethics for the Christian should be based solely on the Bible. There are many other conceptions and details related to these debates that could be discussed, but those are debates for another thesis.

[2] ‘The word deontology derives from the Greek words for duty (*deon*) and science (or study) of (*logos*). In contemporary moral philosophy, deontology is one of those kinds of normative theories regarding which choices are morally required, forbidden, or permitted. In other words, deontology falls within the domain of moral theories that guide and assess our choices of what we ought to do (deontic theories), in contrast to those that guide and assess what kind of person we are and should be (aretaic [virtue] theories). And within the domain of moral theories that assess our choices, deontologists—those who subscribe to deontological theories of morality—stand in opposition to *consequentialists*’ (Alexander and Moore, 2020, NP).

Natural Law is a form of deontological ethics.

In many ways Bonhoeffer does not fit nicely into any of these differing conceptions of Christian ethics. He is critical of the tendency within Christianity towards 'systematic thinking' (Harasta, 2014, p14); particularly when it comes to ethics. On the one hand Bonhoeffer appears to link Christian ethics to Christ and to say that the source, norm, aim, and goal of Christian ethics is Jesus. On the other hand, while he implies that it is the Christian who is responsible, he also appears to argue that 'nobody can altogether escape responsibility, which means vicarious representative action' (Bonhoeffer, 1949/2009, p 258) which we will discuss in more depth later[3].

Bonhoeffer does not fit neatly within the current or past frameworks of Christian ethics. However, what is clear is that he recognises that things as they were, at the time he was writing, were not right and that change was needed to combat the horrors of Nazism and the paralysis faced by the Church which was at the time dominated by a harmful conception of the Two Kingdoms.

However, it is significant and important to recognise that Bonhoeffer's conception of responsibility rejects a utilitarian or consequentialist conception of ethics where the consideration of ethical action is purely based on the consequences (Bonhoeffer, 1949/2009, p267). For example, Bonhoeffer would not agree with the Utilitarian conception that the moral or good act is an act that causes the greatest pleasure or greatest good for the greatest number. Unlike a utilitarian or consequentialist, who considers primarily the consequences rather than other factors, Bonhoeffer would argue that one cannot simply look at the consequence but that we must also consider intention, rules, norms, and principles, but then to go beyond these categorisations and to even take responsibility for that, which at first glance, may not conceptually be my direct responsibility. Likewise, Bonhoeffer's ethics goes beyond that of the situational ethicist because of the element of the call and command of God – the context of responsible ethics must always go beyond a mere assessment of ethics within the present situation. As Matthew D Kirkpatrick said 'Bonhoeffer was clearly concerned with the situation and context of Christian action… however, Bonhoeffer combines these clear emphases with his understanding of a God of creative order, who not only desires human maturation and fulfilment, but created it. The situational and structural elements of Bonhoeffer's ethics limit one another' (2016, p112)

[3] Bonhoeffer here appears to be implying that Responsibility could apply to non-Christians as well as Christians.

It could seem to some as though Bonhoeffer is Utilitarian in his outlook on ethics, but this couldn't be further from the truth. Bonhoeffer proposes that we seek the *Real* - which means to see the reality of the world/ the situation at hand as it is but in the light of the incarnation of Jesus Christ. Bonhoeffer's 'interpretational ethic is neither an absolute unleashing of individual liberty nor a desperate clinging to an old moral system, but an interpersonal life guided by a centre that reaches out to others' (Arnett, 2005, p219). While it is true that Bonhoeffer is concerned with the 'situation and context of Christian action' (Kirkpatrick, 2016, p112) it is also true that Bonhoeffer, unlike situationists and consequentialists, also speaks very clearly about the existence and importance of 'God's command' (Kirkpatrick, 2016, p112) as playing a part in ethics. Thus, neither individualism, consequentialism, nor deontology.

Instead of a utilitarian/ consequentialist method of ethics (which is solely focused upon the consequences and outcomes of an ethical action), Bonhoeffer is arguing for an ethics that can 'consider reality and do what is necessary' (1949/2009, p268): an ethic that is not a slave to norm, intention, consequence, or content, but a system that 'seeks to understand the entire given reality in its origin, essence and goal' (Bonhoeffer, 1949/2009, p267-268). As Esther Reed points out: 'unlike dominant Western models of responsibility that move from agent to act to consequence, Bonhoeffer's dynamic of responsibility moves in the other direction. 'Because of Jesus Christ, and *only* [4] because of Jesus Christ, ***responsibility is defined in terms of the concrete call of the other***[5]' (2020, p89). As a result, responsible action is not about me or myself in any way shape or form; it is always and only about the other. Responsibility places the other person first. In many ways to be responsible means to imitate Christ and to imitate Christ implies self-sacrificial love. As a result of the Christological context of Bonhoeffer's ethics of responsibility we are no longer the subject of ethics as individuals but as a collective: as the Body of Christ and as the *Sanctorum Communio* (1930).

We have seen that Bonhoeffer rejects individualism but also rejects strict deontological legalism and how *the other* is key to Bonhoeffer's mediation between deontology and consequentialism. For Bonhoeffer, all ethics must come back to Christ, and therefore he presents the Christological link between law and that which is real which we respond to through vicarious responsible action as found within the conscience. Conscience, for Bonhoeffer, is a

[4] Authors emphasis
[5] My emphasis

place in which the Christian encounters Jesus Christ; just as he is encountered in the law. As Bonhoeffer argues; for the Christian 'Jesus Christ has become my conscience' (Bonhoeffer, 1949/2009, p279). The link point between law and vicarious responsible action is that within the conscience of the Christian- Christ is present. The law is written on the heart of the believer and in the interplay between law and gospel: the Christological context of conscience allows for vicarious responsible action. Bonhoeffer does not argue that an action that violates an external deontological law is necessarily always a failure of responsibility. However, he does argue that 'responsible action that would violate one's conscience…would indeed be reprehensible' (Bonhoeffer, 1949/2009, p277). At the same time- one challenge for this study, in its considerations of a context wider than Christianity in the form of AI ethics, is that Bonhoeffer is very clear that there is a difference between the conscience of a justified Christian and that of the non-Christian (Bonhoeffer, 1949/2009, pp275-278).

Bonhoeffer argues that vicarious responsible action is possible for the Christian precisely because the conscience that has been redeemed in Christ has been 'set free from the law' (1949/2009, p279) and is now capable of aligning itself 'with the responsibility, which has been established in Christ, to bear guilt for the sake of the neighbour' (Bonhoeffer, 1949/2009, p279). Therefore, within conscience there is reconciliation between law, norm, and responsibility. Conscience allows Bonhoeffer to make Jesus Christ the norm of Christian ethics because, he argues that, within the Christian conscience Christ is encountered and vicarious responsible action is possible. At the same time conscience allows Bonhoeffer to be clear that he is not saying that we reject out of hand the demands of the law or the commandments (1949/2009, p282). Ultimately Bonhoeffer argues that 'responsibility is bound by conscience, but conscience is set free by responsibility… those who act responsibly become guilty without sin; and only those whose conscience is free can bear responsibility' (1949/2009, p282). True conscience is never individualistic. Instead, 'those who in acting responsibly take on guilt- which is inescapable for any responsible person- place guilt on themselves, not on someone else; they stand up for it and take responsibility for it' (Bonhoeffer, 1949/2009, 282).

As we can see a key feature of Bonhoeffer's conception of responsibility is a move away from an individual focused ethic and a move towards an ethics of the community and of the common good. Indeed, one must question what the individual is without the context of their community. We cannot divorce ourselves from our childhood, our employment, our families, and those around us. Likewise, we cannot act as an isolated individual when it comes to our moral decisions- especially in cases that are complex and affect many individuals - such as climate change or more significantly for this paper - Artificial Intelligence.

One complexity that we face when considering Bonhoeffer's conception of responsibility is the question of who is responsible? This question centres around the identity of the '*responsible agent'*. While Bonhoeffer does, on the one side, seem to suggest that responsibility (especially vicarious responsible action) is a response, in conscience, to the revelation and incarnation of God as seen in the person of Jesus Christ: leading to only Christians being able to be responsible agents… On the other hand, Bonhoeffer also seems to argue that, to a degree, all persons share in this responsibility (it should be added that this is only, also, by virtue of the incarnation). This is because Bonhoeffer is situated within his Christological context, and this is the context on which he builds his argument. Yet, it is important to note that for Bonhoeffer this Christological context leads to him arguing that, unlike some Christian schemes of ethics, because of the incarnation and resurrection of Jesus Christ everything changes. At the same time, this reality is not an escape from 'the tasks and challenges of this world' (Harvey, 2015, p3) but is instead about seeing and encountering that which is *real* in a new way. Bonhoeffer's vision of responsibility is not about some idealised society but about being present to the here and now. Bonhoeffer believed that we are called to action - to make the kingdom of God present here and now, and this is essential for what he means by saying that we need to act responsibly.

This Christological context of Bonhoeffer is deeply connected to the questions of the 'relationship between the universal and specific identity of Christian social ethics' (Nissen, 2011, p321). Bonhoeffer's conception of ethics is different to many others in that he argues that there is a specific and unique Christian Social ethics. At the same time Bonhoeffer's ethics does not recognise the idea of different 'realms [Räume]' (Nissen, 2011, p339) because 'human beings as whole persons partake in the one reality of Jesus Christ' (Nissen, 2011, p339). This complexity seems to suggest that there is a 'unity and difference of the universal and specific identity of Christian social ethics at the same time' (Nissen, 2011, p323) in Bonhoeffer's theology; and that this could contribute to the confusion surrounding how one is responsible and who is responsible. Much of Bonhoeffer's ethics leans into this relationship and complexity especially given Bonhoeffer's conception of the 'profound this worldliness of Christianity' (Harvey, 2015, p3). This vision of Christianity does not, as we have already observed, see a distinct realm of the Church and a realm of the political life but sees the role of the Church as making *real* the mystery that all has been and will be reconciled in Christ (Nissen, 2011, p340). Bonhoeffer is deliberately rejecting any concept that faith 'occupies a sphere of operations utterly separate from the "worldly" concerns of politics, economics, and the like' (Harvey, 2015, p20) as in many ways this concept itself led to the moral problems that arose during the Nazi period in Germany. Instead, Bonhoeffer sees this world and the reality made present in Christ as being inextricably linked, because of the incarnation, and therefore our responsibility is located in Christ and because of Christ cannot be divorced from our civil lives outside the Church.

With good reason, Bonhoeffer argues that:

> 'There are occasions when, in the course of historical life, the strict observance of the explicate law of a state, a corporation, a family, but also of a scientific discovery, entails a clash with the basic necessities of human life. In such cases, appropriate responsible action departs from the domain governed by laws and principles, from the normal and regular, and instead is confronted with the extraordinary situation necessities that are beyond any possible regulation by law.' (Bonhoeffer, 1949/2009, p272-273).

Bonhoeffer was situated within the context of Nazi Germany where one would be required to consider what was the right thing to do apart from the law of the state precisely because the law of the state was not moral. We are not, currently, situated within such a dire situation. However, we can face in situations of scientific discovery (such as with AI) the opposite problem - a lack of governance or law at all. Responsible action, according to Bonhoeffer, requires the moral agent to act not simply out of obedience to a rule, norm, or principle but out of a recognition of the urgency of the situation at hand in response to the *real*. Responsible action requires the agent to recognise that rules and principles may not always be correct, or relevant to the current ethical situation, and that people can make mistakes. Yet responsible action calls for everyone to recognise this and to take responsibility for themselves and for their actions - and even for the actions of others at times.

Bonhoeffer's concept of responsibility 'involves the interplay of persons, story, and historical moment; this interplay unites faith and creative application in the concrete historical situation' (Arnett, 2005, p8). Bonhoeffer and his ethics of responsibility is deeply rooted in the current moment and is centred upon the belief that Christianity is a faith that centres upon that which is *real*. This concept of the *real* recognises the reality of the situation within which one is situated; but also recognises the greater picture and, for Bonhoeffer, the reality of the revelation of God in Jesus Christ.

As Arnett states ‘Bonhoeffer, of World War II Nazi Germany, engaged his moment; he would do likewise today’ (2005, p9). Bonhoeffer is focused on the present - on the here and on the now. Ethics, for Bonhoeffer, goes beyond legalism, or emotivism, or consequences, but it instead goes towards a wholistic ethics which recognises the real, as revealed in the life, death, and resurrection of Jesus Christ, and acts according to that acknowledgement of the real. Yet, the very recognition of the *real*, for example with climate change, also recognises that we cannot do it (make change) on our own. We cannot change the world on our own. However, one who acts responsibly can be a voice for change and the one who trusts in Christ to do in us that which we cannot do on our own (as with a Lutheran conception of salvation- so too with our ethics) has hope for the future. Morality as responsibility calls us to go beyond law and principles but also to internalise the moral law, principles, etc but then for the individual to apply law, norm etc, *within their own context*. Responsibility recognises the reality of the situation and responds to it.

In response to the call of responsibility in Bonhoeffer one finds the concept of *vicarious responsible action.* Vicarious responsible action is in many ways central to how we can go forward with Bonhoeffer’s conception of responsibility. When one regards oneself as responsible for the other, and not simply for oneself, each individual acts differently. To take on guilt and to take on responsibility for the other when you do not have to in many ways gives us a path beyond law and towards a Christocentric ethics focused on the Beatitudes. In Matthew 5:20 Jesus says:

‘For I tell you that unless your righteousness surpasses that of the Pharisees and the teachers of the law, you will certainly not enter the kingdom of heaven’ (NIV).

Bonhoeffer is not calling us to ignore law and what was expected in the old covenant, or Natural Moral Law - he is calling us to go beyond them and to apply what we learn from Christ within the situation in which we find ourselves rather than being legalistic. He is arguing that we are called to be real and respond in a way that considers everything - all of existence and the experience of the self.

The question remains, however, as to how this can possibly be practically applied today. Without intending to compromise or contradict Bonhoeffer's notion that ethics needs to go beyond principles or clear boundaries there several key themes that do emerge:

1) **Responsible Ethics rejects the division of Church and State as distinct ethical topics**. It is imperative that the Church and religious voices be able to reject and resist the law or command of the state (and vice versa). However, Bonhoeffer is not saying that he disagrees with the division of Church and State - what he is saying is that **to be responsible can at times require one to speak out against the State (or larger organisation such as a company, even the institutional Church)**. The Church, and by extension the responsible agent in terms of Bonhoeffer's thought, has the right to participate in political action because the Christian does not exist in a separate reality from the political reality but rather the Christian, in particular, is called to make Christ present in the world through responsible action.

2) **To be responsible is to be one who seeks the good of the other. 'Responsibility is defined in terms of the concrete call of the other' (Reed, 2020, p89).** Those who are responsible do not simply see what is before them but look at the wider context and ramifications of their actions. It may be the case that I am not responsible for the poor conditions of a diamond miner in Africa but if I buy a diamond ring, for example, I become responsible for the conditions of that individual because I am called to be responsible for my actions but also to be a voice for the voiceless- to be one who puts the other first. Responsibility requires us to be consumers that recognise the impact of our consumption on others- so too when we provide any goods or services. In all things we must consider the consequences of our actions, not in isolation as the Utilitarian does, but consider them within their context and within the context of history, ethical norms, and the full reality of the situation in which one finds oneself. Especially with the complex situations we face with Artificial Intelligence.

3) **Responsible Ethics as a call to community and to action-** as opposed to individualism. Bonhoeffer is not a liberal in the sense that he does not believe that one should just do what one likes and that there is no ultimate law or moral law that is to be adhered to. The call to community runs throughout Bonhoeffer's writings and is a key theme. This call to community is the antithesis of individualism (which one could argue is the foundation of liberalism). However, based on a Lutheran understanding of

Christianity there is on the other hand an emphasis on the *Freedom of a Christian*. This emphasis on conscience and on freedom might seem to contradict the claim that Bonhoeffer rejects individualism. Yet, it is important to remember that the difference here is intention: we are free *for the other* not for the self. This is a call to action in and through community for the sake of the community. In today's world one could even argue that community in this sense is a global obligation and thus we have an obligation to act for all because we live in a globalised age.

4) **Vicarious Responsible Action**. We are called to act, in conscience, as if we are responsible for people even when we may not be. This goes beyond situations where we are responsible simply out of the concrete call of the other (Reed, 2020, p89) and instead argues that we are to take on the responsibility of others even when we have no requirement to per se. We are to act for the good of the other- even when it may be detrimental to us. On the cross Jesus takes responsibility for our sin- for the sin of the world. Jesus went to death in our place. Likewise, when required, we are to take on responsibility in the place of others for the sake of others. In a practical sense this might mean that when a company or organisation fails in its ethical duty (for example in the duty of safety etc of AI) it is down to the responsible agent to take on the responsibility of this work upon themselves rather than leave it to the company or government. Persons at every level are called to vicarious responsibility when they are failed by those who should be responsible. One cannot escape the call to responsibility simply because one is doing what one is told to do. For example, in Nazi Germany the state was responsible for the holocaust however- vicarious responsibility means that those who carried out the atrocities and those who stood by and did nothing are also responsible. Bonhoeffer acknowledges that vicarious responsible action is risky- it is potentially costly. However, for the sake of Christ, we are called to accept this risk for the sake and for the good of the other.

5) **One who is responsible is one who is guilty. Like with Vicarious Responsible Action- Jesus bore our guilt on the cross. 'Those who seek to avoid guilt and maintain their innocence became bystanders' (Harvey, 2015, p287)** to the horrors of Nazism. Sometimes in life ethical choice is not a choice between right and wrong but between good and good, evil and evil. To be responsible can mean doing something that could be seen as wrong or unethical - such as in the case of a dangerous AI machine

leaking sensitive information to the press to expose wrongdoing. However, the responsible agent accepts responsibility for the guilt that they receive because of their actions and throw themselves upon the mercy of God. Essentially this call to accept guilt means that in our response to the situations in which we find ourselves we need to do what is right- irrespective of consequence or law: even if on some levels what we do might be wrong the responsible agent takes this guilt upon themselves so long as it is done for the other rather than for the self.

Hopefully these observations can help us to begin to better understand how to formulate a usable conception of Bonhoeffer's ethics of responsibility.

1.4 Responsibility as Vocation

Bonhoeffer begins his chapter on Responsibility as Vocation by making clear that what he is talking about is not what he calls ‘the secularised concept of vocation as ‘a defined field of activity’ (Max Weber)’ (Bonhoeffer, 1949/2009, p289). Likewise, Bonhoeffer is not saying that vocation in this sense is simply ‘the sanctification of worldly orders’ (Bonhoeffer, 1949/2009, p289). Instead, Bonhoeffer argues that ‘vocation is the place at which one responds to the call of Christ and thus lives responsibly’ (Bonhoeffer, 1949/2009, p291) in a way which engages in the ‘struggle against the world’ (Bonhoeffer, 1949/2009, p291) whilst not giving up and attempting to flee from the world or the reality in which the individual finds themselves. For Bonhoeffer vocation has nothing necessarily to do with work or worldly calling but everything to do with our calling to live responsibly (as always for the sake of Jesus Christ).

The word *vocation* was very prevalent in Bonhoeffer’s time. As we have already seen, in Bonhoeffer’s reference to vocation, Max Weber uses it frequently. As much as Bonhoeffer likes to set himself up against the concepts of Weber- he is clearly influenced by them. For example, while Bonhoeffer would argue that Weber is mistaken in arguing that ‘religious interpretations’ do not do ‘anything to enhance the worth of purely human relationships’ (Weber, 1917/2004, p30). Yet on the other side, it is also clear that Bonhoeffer would agree with Weber’s assertion that ‘the ultimate and most sublime values have withdrawn from public life’ (Weber, 1917/2004, p30). Likewise, seeing the world as it really is, is also a key theme in Weber as it is in the writings of Bonhoeffer. While Bonhoeffer’s conception of vocation goes beyond Weber’s: Weber’s concepts do still play a part in Bonhoeffer’s conception of vocation.

Another key player in Bonhoeffer’s concept of vocation will, naturally, be Martin Luther. Luther’s conception of vocation is broad and something that cannot be dealt with in much detail here. However, what is important is that Luther pointed the conception of vocation away from simply a concept of vocation to the religious life and/ or the priesthood and instead, discussed the fact that vocation is something that we all have. Our vocation is a calling from God to live a faithful life wherever we find ourselves. At the same time Luther often uses the German word “*Beruf”* as being vocation and associated it to ones ‘outer status or occupation’ (Wingren, 2004, p1) much as Weber does. Bonhoeffer also uses this word. Yet his understanding of “*Beruf”,* in the Christian sense, should go one step further and be more than just worldly activity but the sanctification of activity in the world (Wingren, 2004, p2). At the same time ‘Luther does not use *Beruf* or *vocatio* in reference to the work of a non-Christian' for Luther '*Beruf* is the Christian's earthly or spiritual work' (Wingren, 2004, p2). Or as Bonhoffer might put it: vocation is the Christian's response to the present and real situation in which one serves their brothers and sisters; it is a part of the 'mysticism of our ordinary life' (Veith, 2021, p124).

Bonhoeffer's concept of vocation as responsibility importantly also does not seek to make us immune from a sense of sacrifice or of the reality of the need for the Christian to bear the cross. Sometimes being responsible, as a form of vocation, means to do things that you would rather not do. For example, being a whistle-blower at a company involved in the creation of Artificial Intelligence, breaking a confidentiality agreement, even breaking the law, to be a responsible agent that cares for the other takes courage, suffering, and risks a lot. When responsibility is our vocation 'giving up oneself out of love, as Jesus did, becomes the template' (Veith, 2021, p156) for how we are to live responsibly in the reality of our daily lives.

Fundamentally, for Bonhoeffer, vocation is about our response to the call of Jesus Christ. Responsibility as Vocation means that responsibility is not an added extra. Rather it is our vocation, our calling as Christians and is inescapable. It is clear for a Christian, but it is unclear how much this is true for the non-Christian. However, for the Christian, responsibility takes its shape from two sides...our call to responsibility from the other...and from our vocation - or our call - to responsibility from God. The Christian cannot avoid this call. Rather responsibility as vocation situates responsibility at the heart of what it means to live as a Christian in the world. When it comes to Artificial Intelligence - ensuring that things are open, inclusive, and for the betterment of humanity are areas in which responsibility as vocation can be expressed in the development and implementation of AI; as we will see more later.

1.5 Who is responsible?

‘Responsibility begins in in encounter with the other person’ (Reed, 2020, p17).

The *responsible agent* in Bonhoeffer’s work is far from clear. While Bonhoeffer does, on the one side, seem to suggest that responsibility is a response to the revelation and incarnation of God as seen in the person of Jesus Christ: leading to only Christians being able to be responsible agents… He also seems to argue that, to a degree, all persons share in this responsibility (it should be added that this is only, also, by virtue of the incarnation).

On the one hand, Bonhoeffer seems to suggest that only the Christian can be truly responsible. If Jesus is the norm, end, goal, and source of true ethics; it is hard to see how such ethics could in any way be seen as a system that would be available to or understandable by the non-Christian. Bonhoeffer states ‘It is not the isolated individual but the responsible person who is the proper agent to be considered in ethical reflection’ (Bonhoeffer, 1949/2009, p 258). Thus, the question of *who the responsible person is* must remain central to the discussion. It is important to note, as Bonhoeffer points out early in *Ethics* (1949/ 2009), that the topic of his consideration is Christian Ethics (Bonhoeffer, 1949/2009, p 47). Likewise, he also argues that ‘the source of a Christian ethic is not the reality of one’s own self, not the reality of the world, nor is it the reality of norms and values. The beginning and end of our ethical considerations is the reality of God that is revealed in Jesus Christ’ (Bonhoeffer, 1949/2009, p49). As a result, it would seem to suggest that the responsible agent that he mentions is a Christian simply because Christian ethics is the subjects of his text. Indeed, at one point Bonhoeffer even defines true vocational responsibility as being ‘responsibility to Jesus Christ’ (Bonhoeffer, 1949/2009, p293). Likewise, Bonhoeffer identifies conscience as the place where we take upon ourselves and ‘bear guilt’ out of love for our neighbour (Bonhoeffer, 1949/2009, p280) and he goes on to argue that this is possible only in a conscience which is ‘freed in Jesus Christ’ (Bonhoeffer, 1949/2009, p281).

However, Bonhoeffer seems to argue that in one sense ‘Nobody can altogether escape responsibility, which means vicarious representative action’ (Bonhoeffer, 1949/2009, p 258). One prominent and current example would be climate change. We are all responsible, whether Christian or not, because we are human persons, and it will affect us all. Or as Bonhoeffer puts it ‘responsibility for myself is in fact responsibility for human beings as such, that is, for humanity (1949/2009, p 258). Just as original sin impacts all human persons simply by virtue of being part of humanity (both Christian and non-Christian), likewise, our good and bad actions impact all of humanity as nothing is ever done in isolation from our existence as a part of the human family. Bonhoeffer argues that ‘Other people who are encountered must be regarded as responsible as well’ (Bonhoeffer, 1949/2009, p269). Therefore, suggesting that those who may or may not actually be responsible should be regarded *as if* they are – even if one concludes that they are not.

Bonhoeffer is writing *for and to* the Christian Community. However, as Esther Reed points out, Bonhoeffer’s ethic of responsibility does not simply have consequences for the Church but also for the whole of humanity (2020, p101). She goes on to point out that ‘while a Christologically informed ethic of responsibility can be understood only from the context of the Church-community... it does not follow there are no implications for a Christian... ethic of responsibility that speaks to non-believers as well as believers’ (Reed, 2020, p101-102). At the end of the day Bonhoeffer’s ethics of responsibility is about the rejection of priority of self and to shift to the embracing of responsibility for all persons as a way of life which provides solutions to issues faced in this modern age. Likewise, it is also important to note that the book Ethics was never finished (Green, 2009, p1), and Bonhoeffer never finalised his text - due to him being executed before it could be completed - and as a result he might well have intended for this to be where his conception of responsibility would have ended up. In many ways one could argue that this conception of responsibility is Bonhoeffer’s response to the lack of responsibility on the part of Christians in Nazi Germany and his incomprehension at the ways in which the Reich Church surrendered to the power and influence of national socialism.

It is likewise important to remember, as stated earlier, that ‘Bonhoeffer’s work predates the emancipation and expansion of ethics’ (Harasta, 2014, p14) and during his time period ‘the primary basis for ethical claims remained dogmatic theology’ (Harasta, 2014, p15). Therefore, ethics was not seen as a distinct discipline at the time as it is often viewed today. Rather he is writing out of a dogmatic framework that was required of those who were Christians within the discipline of ethics. However, that does not mean that we cannot in some way apply his ethics beyond this framework.

This concept of the non-Christian being responsible is justified in many ways because Lutheran ethics (and thus the context from which Bonhoeffer is writing), as previously mentioned, is not about salvation or grace. Salvation is by Faith Alone (sola fide) through Grace Alone (sola gratia), to a Lutheran, and therefore ethics is about how to live well in the world, with others, and with God; and not about anything specifically to do with salvation or grace. The Christian, as one who has been justified and is being sanctified by the means of grace, could be seen as being more responsible or having more of an inherent duty of responsibility: because to be responsible is to live as Jesus lived and to go beyond the mere religious observance of the law. However, as this has nothing to do with grace, faith, or salvation there is no reason to necessarily limit responsibility to Christians alone even if one is writing to the Christian Community, and from that context, as Bonhoeffer is precisely because responsibility always begins in encounter with the other person. As a result, we can use a Bonhoeffer informed ethics of responsibility to assist in our exploration of the ethics of Artificial Intelligence implementation and development.

1.6 Bonhoeffer on Technology

Dietrich Bonhoeffer does not say much about technology. He certainly does not comment on Artificial Intelligence. However, in his chapter *Heritage and Decay* in *Ethics* (1949/2009) Bonhoeffer does discuss technology briefly.

Under the context of secularisation and the French Revolution Bonhoeffer speaks of 'the cult of radio' (Bonhoeffer, 1949/2009, p115) [6]. Bonhoeffer is not completely negative about radio (or technology in general). He recognises that there are many benefits and gifts that come with technological advancement. However, Bonhoeffer also notes and warns that it is easy for technology to become 'an end in itself' (Bonhoeffer, 1949/2009, p116). He argues that there is a risk that we seek in technology ways to become like a god who thinks that they know better than nature and who attempts to subjugate nature and destroy it. Bonhoeffer argues that it is easy for technology to be used as a tool of nationalism and to lead to destruction and oppression (Bonhoeffer, 1949/2009, p 117-122) as seen in his day.

Bonhoeffer recognises that technology (in this instance radio) can create an atmosphere of 'truthfulness, light and clarity' (1949/2009, p115) with an increase in 'intellectual honesty in all things' (Bonhoeffer, 1949/2009, p115) especially in a society with greater religious and political freedom. Yet he also warns against the 'rise of technology' (Bonhoeffer, 1949/2009, p116) which can lead to the 'violation and exploitation of nature' (Bonhoeffer, 1949/2009, p116).

On the one hand, according to Bonhoeffer, it can be a symptom of a 'naive faith' to protest technological developments as being a challenge to God. He points out that in the 'Islamic world' technology 'remains completely in the service of belief in God and the building of... community' (Bonhoeffer, 1949/2009, p117) whereas it is in the west where we have divorced faith from technology. However, technology in the west is not all bad. Indeed, it is through advances in technology we came ever more to the realisation of the rights of every human person along with a sense of 'common solidarity' (Bonhoeffer, 1949/2009, p117) between all persons.

[6] This appears to be in reference to the revolutionary cult of reason.

Bonhoeffer is in this one side very positive toward the contribution of technological development to the human experience - that through technology we often have a greater chance at freedom and rights. However, Bonhoeffer also states that:

‘The Masses and nationalism are the enemies of reason. Technology and the masses are antinationalistic. Nationalism and technology are the enemies of the masses.’ (1949/2009, p121).

Bonhoeffer warns against the ways in which technology can be used as a tool of suppression and control, as he saw during his lifetime. Likewise, there is a risk that ‘the master of the machine becomes its slave; the machine becomes an enemy of the human being’ (Bonhoeffer, 1949/2009, p122). The creation rebels against its creator just as humanity rebels against God and destroys nature, in some ways a second fall. This argument is echoed and proven true with the use of atomic weapons not long after Bonhoeffer’s death.

It is important to remember that Bonhoeffer is witnessing first-hand the use of radio and cinema as tools of propaganda for the Nazi state. Bonhoeffer recognises this danger and the influence that technology can have over people's lives. We do not have to look far for his justification on this matter. If we look at the way that phones, computers, online shopping, and social media have taken over and dominated our lives over the last few years, clearly these things do possess a strong influence over us. As we can see - technology does have a clear and present influence over our lives. Bonhoeffer’s main concern on this front is the fact that we do not lose our humanity amongst the challenges of technology. He recognises the risks posed by technology to the state and social order but likewise the risk to the freedom of the individual that technology can present.

As Rasmussen points out in his 2009 article on Bonhoeffer and his scientist brother Karl Friedrich Bonhoeffer - ‘it is important to state that science does not do science; scientists and society do’ (Rasmussen, 2009, p98). Both Bonhoeffer and his brother were deeply aware of the need for people to take on responsibility for the state of science and religion during the Second World War. Both Bonhoeffer and his brother were equally aware of the importance of stressing the place of ethics in scientific and technological research and discovery. Bonhoeffer in his dialogues on technology and science, along with his dialogue on the ‘real’, recognises that far too often ‘religion becomes a separated sector of life, a sector of the unknown and inexplicable’ (Rasmussen, 2009, p110). Bonhoeffer sees religion as often being used as peoples last resort in life when in reality religion should be leading us to always regard ourselves ‘as irrevocably accountable’ (Rasmussen, 2009, p111) and responsible. Our growth in technology and science - rather than being a disastrous intrusion on the territory of the divine should instead be viewed as a place where we have the opportunity to act as free and responsible agents - seeking the coming of the Kingdom of God in the here and now. Too often we are tempted to remove ethics from scientific and technological development or leave it as a subsection. We should instead follow the example of Bonhoeffer and ensure that ethics is weaved into the very fabric of our science and technology and made real - as Bonhoeffer argues we should do with God in our day-to-day life.

In our search for responsibility in Artificial Intelligence, we are to seek to become a people who recognise that ‘technical communicators facing the need for invention also need an ethical education’ (Boedy, 2017, p116) which helps them to recognise ethical risks and how to discern them independently – so that they can become responsible agents and recognise their role in seeking the common good. It is easy to justify things when working solely using principles. When one uses a concept that one is truly personally responsible for the consequences of their actions - ethics becomes more than words or tick boxes and instead becomes a way of life. This is ultimately what Bonhoeffer’s ethics of responsibility and technology (and by extension ethics/responsibility of and for AI) is calling us to - a way of life rather than a concept, or idea, or tick box. Bonhoeffer in his conceptions of technology and responsibility calls us back to the ‘real’ and to personal responsibility for the other.

1.7 Other Concepts of Responsibility

Bonhoeffer was not alone in considering the ethical concept of responsibility. There were, and still are, others who have differing interpretations on the topic of what it means to be ethically responsible. With many of them there is a great deal of overlap. However, there are also important differences. Before concluding our onward approach, it is worth considering what others think and using these other approaches to understand how we should move forward and use Bonhoeffer's ethical conception of responsibility in the modern world.

1.7.1 Reinhold Niebuhr (1892- 1971)

Reinhold Niebuhr was a contemporary of Bonhoeffer's and developed many conceptions of political and social ethics - including a discussion on the topic of responsibility. He was a well-known and respected theologian. Bonhoeffer and Niebuhr have several similarities, however, unlike Bonhoeffer Niebuhr survives the war, and did not live in Germany during the war, and this does lead to different understandings of responsibility (Lovin, 2016, p67). Bonhoeffer and Niebuhr did know one another (Lovin, 2016, p68) and therefore the similarities and differences in their conceptions of responsibility are important to consider.

Much like Bonhoeffer, Niebuhr looks back, particularly in his book *Moral Man and Immoral Society* (1934), to the past and analyses the themes and progression of ethical thought as a lived practice: especially in the context of the modern state.

Niebuhr states that in modern society:

'Political power has been made responsible, but economic power has become irresponsible in society. The net result is that political power has been made more responsible to economic power' (1934/ ND, p15).

In essence Niebuhr is arguing that financial issues and the power of industries and corporations can often run the risk of placing the quest for profits above genuine responsibility. This is certainly a key factor that needs to be recognised as we consider an ethics of responsibility for Artificial Intelligence. The question of how we can ensure that responsibility is more than simply a quest for profits and genuine search for moral responsibility and for the welfare of all is going to be a great challenge within modern society. However, this very question is critical to our discussion of the ethics of AI.

Similarly, in Chapter 2 of *Moral Man and Immoral Society* (1934) Niebuhr argues that to think that 'injustice could be overcome by increasing the intelligence of man' (1934/ ND, p23) is unrealistic. Likewise, today when we consider the context of Artificial Intelligence there may be a temptation to consider that this new technology or intelligence will solve all our problems and injustices- when, in reality, it runs the risk of increasing them. It is important in our ongoing discussion of this topic to not fall into this pitfall.

As with Bonhoeffer, Niebuhr argues that the difference between a purely secular ethics and a religious ethic is the preferential option for the other. Niebuhr argues that 'a rational ethic seeks to bring the needs of others into equal consideration with those of the self' whereas 'the religious ethic… insists that the needs of the neighbour be met' (1934/ ND, p57). This preference for the other and recognition that the other should be the focus of our ethics places Niebuhr and Bonhoeffer's ethics in conflict with a capitalistic conception of those businesses where profit is prioritised. However, Bonhoeffer's and Niebuhr's form of ethics allows for the betterment of society and the wellbeing of the other. The real question in this context, when related to Artificial Intelligence, is recognising that ethical AI might entail the loss of profits. One must wonder if this would be agreeable to the multimillion-dollar conglomerates typically involved in the development of such technology.

Rather unhelpfully, just as with Bonhoeffer, Niebuhr died before he could finish the book in which he was pulling together his thoughts on Christian Responsibility. Yet after his death *The Responsible Self* (1963) was published. As James M Gustafson says in his introduction to *The Responsible Self* (1963): 'The Responsible self is an integrating and persistent theme in the ethical thought and teaching of H. Richard Niebuhr. In any presentation of his lifelong reflections that he would have made, it would necessarily be central' (1963, p6). Gustafson sees Niebuhr's conception of ethics as being 'the effort of the Christian community to criticize its moral action by means of reflection' (1963, p8).

In *The Responsible Self* (1963), Niebuhr is arguing for an ethics that, like Bonhoeffer, 'lies beyond the border' (1963, p56) of rules, laws, principles, and norms. For Niebuhr responsibility must come to the very heart of what it means to be human and how we understand ourselves.

Like Bonhoeffer, Niebuhr argues that to be responsible is to be an agent who 'accepts consequences in the form of reactions and looks forward in a present deed to the continued interaction' (1964, p64). However, whereas Bonhoeffer makes the person and work of Jesus Christ the centre of his ethics of responsibility, Niebuhr is keen to point to Jesus as being the example *par excellence* rather than being the one and only means by which one can accept responsibility. However, at the same time, Niebuhr is calling us to ask the question 'to whom or what am I responsible' (1964, p68). Likewise, Niebuhr moves away from considering responsibility as only being applicable to the Christian; Niebuhr 'has no interest in a moral orientation that sets the Christian apart from the world' (Lovin, 2016, p72) instead he is focused on Christian ethics as the task of discovering 'what discipleship means' (Lovin, 2016, p72).

Again, just like Bonhoeffer, Niebuhr also argues that 'responsibility always involves guilt' (Lovin, 2016, p81). Likewise, they agree that a moral action cannot be comprehensively justified either by its principles or its goals' (Lovin, 2016, p82) but that responsibility must be 'the free choice of a responsible agent, acting on behalf of a larger community than self-interest alone provides' (Lovin, 2016, p82).

Yet for Bonhoeffer, responsibility is taken as the exception - not the norm. Bonhoeffer would argue that normally, individuals should not need to act in a way other than those laid out in norms and principles. On the other hand, for Niebuhr, responsibility is a constant throughout our moral lives. In this instance Niebuhr's conception is more likely to be useful. This is because ethics has been so transformed by the technological revolution that we are in a wholly new situation that has no historical precedent and therefore there is a need for our ethics, in this thesis, of responsibility to follow the example of Niebuhr and be the norm rather than the exception.

Significantly, for Niebuhr, responsibility is something that we all bear together and not as an isolated agent, as Bonhoeffer suggests (Lovin, 2016, p84). This is significant in emphasising the role of the state and political institutions - and all citizens - as playing a role in responsible action. This development of Bonhoeffer's thought allows for us to move beyond simply individual action and towards collective political action which will be important in shaping the future of our civilisation during the development and implementation of AI.

1.7.2 Paul Ricoeur (1913-2005)

'Paul Ricoeur was one of the leading French Philosophers of the twentieth century' (Gregor, 2016, p259). He was influential in concepts of Christian ethics in the debates around whether there is a distinctive Christian ethics. ‘Ricoeur reportedly met Bornhoeffer while visiting Germany in the 1930’s’ and ‘on several occasions Ricoeur cites Bonhoeffer as a model for the sort of Christian faith that Ricoeur envisions beyond the atheistic critique of religion’ (Gregor, 2016, p260). His philosophy is useful in our discussion because, while being a contemporary and enthusiast of Bonhoeffer, he suggests that to promote true responsibility we must move our sense of responsibility beyond simply the Christian and to responsibility because of a shared humanity. Ricoeur is, like Bonhoeffer, a Christian, however his stress on the need for legal responsibility and responsibility as a shared human experience allows our considerations to extend beyond those who are Christians.

In Ricoeurs philosophy ‘we are told to love in freedom and to take responsibility for the other, even if we do not know him/her, and they are told to do the same for us. We are invited to remember that we exist in relation to ‘others’ and that these ‘others’ are responsible for our freedom in the same way that we are responsible for theirs’ (Mealey, 2009, p75). In many ways this echoes Bonhoeffer but also points towards the fact that this concept of responsibility in ethics also needs to be applied beyond the borders of the Church alone. Likewise, ‘Ricoeur tells us, ‘Loving neighbour as self’ means we are called to take responsibility for the other, and the other is called to take responsibility for us.’ (Mealey, 2009, p75) and that ‘we are called to take responsibility for the freedom and flourishing of others, while others are called to do the same for us’ (Mealey, 2009, p75).

While much of this seems similar to Bonhoeffer it is important to note that ‘Bonhoeffer’s emphasis on the unique personhood of Christ is absent in Ricoeur’ (Gregor, 2016, p281). Ricoeur instead seems to push responsibility away from the individual’s relationship to Christ and towards a responsibility in common for humanity.

Ricoeur’s concern was far more with law and philosophy than simply with ethics and theology. However, in *Reflections on The Just* (2007) he does discuss responsibility in some more detail. He disagrees with Bonhoeffer somewhat as he argues that one can be ‘responsible but not guilty’ (Ricoeur, 2007, p324).

He argues that there are three elements to his working definition of responsibility:

‘1. I hold myself responsible for my acts’

‘2. I am ready to render an account before a body authorized to demand such an account from me’.

‘3. I am in charge of the operations of some private or public institution’ (Ricoeur, 2007, p324).

However, the issue with this, as we have already seen, is that in the instance of moral issues such as Artificial Intelligence: the responsible person will often not fit into one of these categories and will regularly have to be someone who is not in charge of an organisation, or able to necessarily render account to an authorized authority (Ricoeur, 2007, p324) for their actions as he suggests.

However, Ricoeur’s attempt to apply an ethics of responsibility to a philosophical and legal framework is significant. If our ethics are to have any impact, they must be able to be applied in the world around us. Our moral systems cannot remain mere ideas or else we will have failed in that which we have set out to achieve. Widening our ethics of responsibility beyond ‘Bonhoeffer’s emphasis on the unique personhood of Christ’ (Gregor, 2016, p281) is of importance to the applicability of an ethics of responsibility to the topic of artificial intelligence technology.

1.7.3 Hans Jonas (1903-1993)

In his book *The Imperative of Responsibility: In Search of an Ethics for the Technological Age* (1984), Hans Jonas specifically discusses the relationship between modern technological development and the concept of responsibility in ethics. He argues that there is a need for this new form of ethics because ‘our collective technological practice constitutes a new kind of human action’ (Jonas, 1984, p23). Due to the nature of technological change, Jonas argues, the old premises of ethics ‘no longer hold’ (1984, p1). As Jonas says:

‘Modern technology has introduced actions of such novel scale, objects, and consequences that the framework of former ethics can no longer contain them’ (1984, p6)

Jonas is a useful figure to consider in this paper because of his context. He comes from a similar timeframe to Bonhoeffer and yet he is not a Christian (he was Jewish). He, like Bonhoeffer, was also German and involved in opposition to Hitler in various ways. Jonas believed that there was a need for a new ethics in the modern world and gives us the possibility of applying responsibility beyond a merely Christian imperative and to a human imperative.

Likewise, it is notable that he recognises that there is a need for human persons to consider what it means to be responsible in a technological age.

Jonas uses the example of climate change to demonstrate the changes in the way that, in the context of more complex and globalised modern problems, we should alter how we see ethics (1984, p6-7). The framework of our ethical discussion is no longer focussed on narrow actions and consequences but far wider reaching than ever before and therefore there is a need for us to look at things in a new way. Precisely because man has ‘become dangerous not only to himself but to the whole biosphere’ (Jonas, 1984, p136), Jonas argues, we must move beyond ‘self-interest’ (Jonas, 1984, p136) and take on responsibility for ourselves, the consequences of our actions, and one another as fellow human beings.

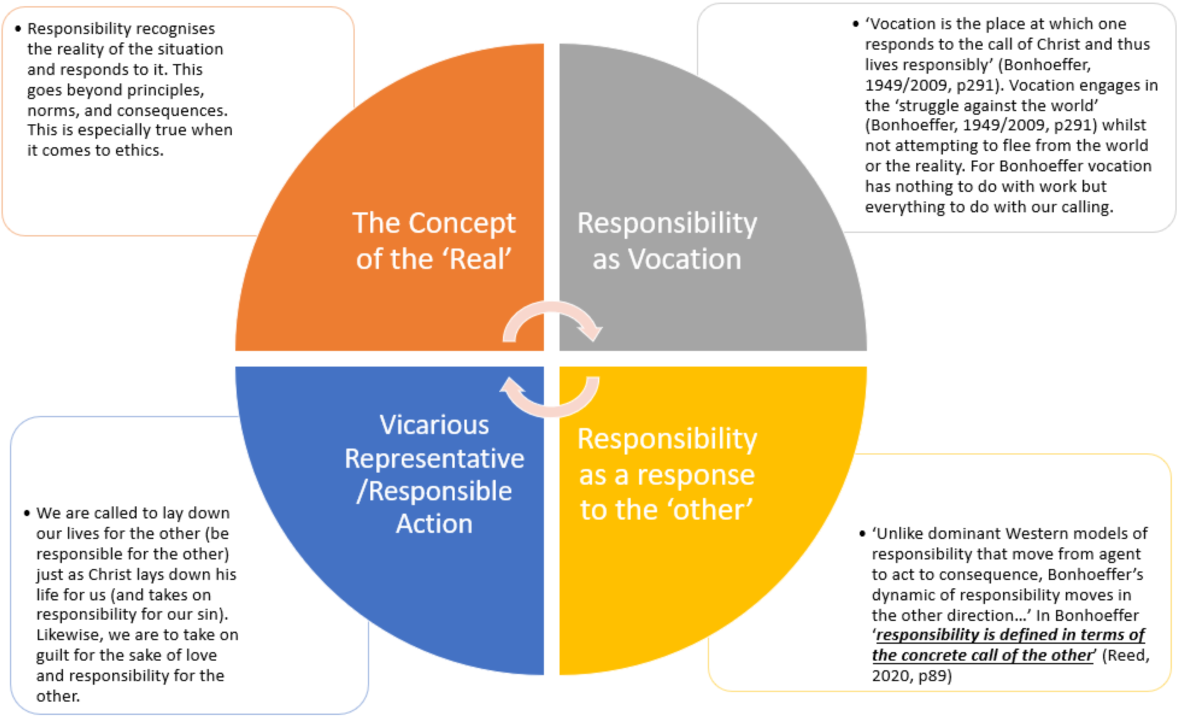
In Chapter 4 of *The Imperative of Responsibility: In Search of an Ethics for the Technological Age* (1984) Jonas argues that ‘the first and most general condition of responsibility is casual power’ (1984, p90) meaning that even if it is minimal, we do have and can have power to influence things (such as in what we buy or promote). He argues that even if we do not directly cause something to happen - the fact that we can do something about it and can recognise that in some way it is a consequence of our actions- than we are responsible.

In essence Jonas has recognised the argument that this new and modern technology demands of us a new and modern ethics that can recognise the greatly enhanced range of consequences that are the result of technology. However, the difference between Bonhoeffer and Jonas is the concept of vicarious responsible action and the person of Jesus Christ as the cause of our responsible action. Jonas, as someone who is not a Christian, does make responsibility more accessible to those outside of the Christian faith but fails in many regards to really answer who or what we are responsible for. Bonhoeffer’s concept of vicarious responsible action, arguably, goes one step beyond Jonas in demanding that the ethical decision maker not only take on responsibility for the consequences of their actions but also take on the guilt of said consequences.

There are many other theologians and philosophers who could have been considered. However, the three that have been discussed are significant given the time in which they wrote and the relationship between the subject that they discussed and the subject as discussed in Bonhoeffer.

1.8 Conclusion

The elements of Bonhoeffer’s concept of Responsibility are as follows:



- Responsibility recognises the reality of the situation and responds to it. This goes beyond principles, norms, and consequences. This is especially true when it comes to ethics.

The Concept of the ‘Real’

Responsibility as Vocation

- ‘Vocation is the place at which one responds to the call of Christ and thus lives responsibly’ (Bonhoeffer, 1949/2009, p291). Vocation engages in the ‘struggle against the world’ (Bonhoeffer, 1949/2009, p291) whilst not attempting to flee from the world or the reality. For Bonhoeffer vocation has nothing to do with work but everything to do with our calling.

Vicarious Representative /Responsible Action

- We are called to lay down our lives for the other (be responsible for the other) just as Christ lays down his life for us (and takes on responsibility for our sin). Likewise, we are to take on guilt for the sake of love and responsibility for the other.

Responsibility as a response to the ‘other’

- ‘Unlike dominant Western models of responsibility that move from agent to act to consequence, Bonhoeffer’s dynamic of responsibility moves in the other direction…’ In Bonhoeffer ‘***responsibility is defined in terms of the concrete call of the other***’ (Reed, 2020, p89)

Figure 1:The elements of Bonhoeffer’s concept of Responsibility[7]

In response to these concepts, we can develop the following ideas of what it means to be ethically responsible, for the purpose of this thesis.

1) Responsible Ethics rejects the division of Church and State as distinct ethical topics. To be responsible can at times require one to speak out against the State (or larger organisation such as a company, even the Church etc).

2) To be responsible, in ethics, is to be one who always seeks the good of the other. ‘Responsibility is defined in terms of the concrete call of the other’ (Reed, 2020, p89).

3) Responsible Ethics as a call to community and to action.

4) Vicarious responsible action is needed to act in a way that really is ethically responsible.

[7] Figure 1: Taken from *What can we learn from Dietrich Bonhoeffer’s Concept of Responsible Action* (Nelson, 2022 A, NP)

5) One who is ethically responsible is one who accepts guilt. Like with vicarious responsible action- Jesus bore our guilt on the cross. ‘Those who seek to avoid guilt and maintain their innocence became bystanders’ (Harvey, 2015, p287) – as many in the Church did during Bonhoeffer’s own day.

Therefore, according to Bonhoeffer, for one to be responsible one must act in these ways:

1) Selflessly

2) In the interest of others (seek the Common Good)

3) In a way that takes on guilt- is willing to be responsible (vicarious responsible action)

4) Is prepared to act.

Most of the authors that we have considered would, to some degree or another, agree with these general points. However, what we can gain in addition from the extra authors that we have discussed in this chapter is that:

- Responsibility must be universalised as AI is not just a Christian issue but a human issue (Jonas),

- That responsibility must become the norm of our ethical action rather than the exception (Niebuhr),

- And that responsibility can be expressed in our common humanity and not just in relationship to Christ (Ricoeur).

These considerations can help us as we move forward and attempt to apply an ethics of responsibility to the topic of Artificial Intelligence technology.

Morality as responsibility calls us to go beyond law and principle and in response to internalise the moral law and then to apply it in a way that takes responsibility for, not only oneself, but for others. We are all to be responsible and it is in this context that we must consider how to approach an ethics of Artificial Intelligence technology.

Chapter 2: What is Artificial Intelligence and how does it work?

In 1950 Alan Turing wrote a significant article about *computing machinery and intelligence*. In this seminal work he explored what it would mean for a machine to think, and to learn. He defined the subject as a machine that is 'intended to carry out any operations which could be done by a human computer' (Turing, 1950, p436). Turing points out that 'the idea of a learning machine may appear paradoxical to some readers' (1950, p459), just as is true in many ways today. However, one of the major traits that Turing points to as being central to the concept of a thinking/learning machine is that 'its teacher will often be very largely ignorant of quite what is going on inside' (1950, p459). This may seem outlandish to us. However, this is exactly what sometimes happens today in some forms of Artificial Intelligence. As Tom Taulli pointed out in a 2020 *Forbes* article about Artificial Intelligence: 'the neural network learns on its own what is 'good' or 'bad' (2020, NP). A major issue with Artificial Intelligence, and in particular, deep learning today is that often it is 'essentially a "black box". This means it can be nearly impossible to understand how the model really works' (Taulli, 2020).

Turing recommends that the way to train Artificial Intelligence 'could follow the normal teaching of a child' (1950, p460) in that you would need to point out (label) different items to train the machine to recognise different data points, objects etc and then be able to use them (Turing, 1950, p460).

Prior to Turing, in 1942, Isaac Asimov 'publishes his short story *Runaround*'' (Haenlein and Kaplan, 2019, p6). In this book Asimov proposes the '*Three Laws of Robotics*' (Haenlein and Kaplan, 2019, p6) as an early form of ethics for Artificial Intelligence or some other form of thinking machine. Together with this Asimov provided a framework for robotics development that highlighted above all the safety and the protection of human life, even above the safety of the robot themselves (law 3). This work was being written at the same time as Turing was writing his essay on computing machinery and intelligence (Haenlein and Kaplan, 2019, p6) within which Turing sets out how to distinguish if a machine is or is not intelligent by asking a set of questions and looking at how the program responds. The Turing test basically asserts that a machine is intelligent if a person cannot tell if answers given to questions come from a machine or from a person. This test is 'still considered today as a benchmark to identify intelligence of an artificial system' (Haenlein and Kaplan, 2019, p7); even though the 'word Artificial Intelligence was… (only) coined about six years later… in 1956 (by) Marvin Minsky and John McCarthy' (Haenlein and Kaplan, 2019, p7).

As we have seen, Artificial Intelligence has a history dating back for over 50 years and has slowly developed into what it is today. In this time there has been a range of ethical questions and discussions posed and contemplated even dating back to questions and challenges raised by Turing in the 1950s. The biggest risk that we face with Artificial Intelligence is 'a technology that can give you everything you want is a technology that can take away everything that you have' (Schneier, 2018, p217). There are so many issues that need to be addressed from an ethical perspective - as Roszak argued in 1986 - 'the invasion of privacy has been the single most publicized and discussed social issue surrounding computers' (Roszak, 1986, p206) and that was well before the invention of the modern smart phone or surveillance systems that we see deployed today! In the Roman Catholic Church's most recent council- *Vatican II,* it is highlighted that the primary concern of technological development must be the dignity of the human person and that technology must be a tool for human progress, working for our common good, and not a hindrance to the dignity of any individuals or groups within society (GS, 1965, #23-24, 33)[8]. It has of late 'become possible to build far more sophisticated AI systems that can do real work' (Frankish and Ramsey, 2014, p9) even in our day to day lives with examples such as Siri (and other virtual assistants) along with the imminent possibility of working autonomous vehicles. As such, it is essential, currently, to review our ethical approach to this issue and to ensure that everything is done for the good of those for whom this technology is designed to serve and humanity. This new age of Artificial Intelligence is going to present 'unique ethical, legal, and philosophical challenges that… need to be addressed' (Haenlein and Kaplan, 2019, p13) and indeed it already is. As Bonhoeffer already warns in Chapter 1 the 'rise of technology' (Bonhoeffer, 1949/2009, p116) can lead to the 'violation and exploitation of nature' (Bonhoeffer, 1949/2009, p116).

2.1 Artificial Intelligence Terminology

When it comes to Artificial Intelligence there are a wide variety of terms that are often bandied about. To help with our exploration of the ethics of Artificial Intelligence, it is important that we consider what is meant by each of these terms.

[8] GS stands for *Gaudium et Spes*

2.1.1 Algorithms

Algorithms are very important to the development of Artificial Intelligence. An algorithm is a computer program that is essential to the running of any computer or machine. There are a wide variety of algorithms that can be used in the development of Artificial Intelligence models; many more than can be assessed here. Algorithms can carry great risk – as this is a significant stage in which different biases and other issues can begin to arise.

As is often the case when it comes to Artificial Intelligence: the correct use and implementation of data can be more important than having the best or the most effective algorithm (Mueller and Massaron, 2018, p39).

One of the most commonly used form of algorithm is a classification and regression trees (see figure 2 on the next page). These are 'machine-learning methods for constructing prediction models from data' (Loh, 2011, p14). As Wei-Yin Loh explains:

'Classification trees are designed for dependent variables that take a finite number of unordered values, with prediction error measured in terms of misclassification cost. Regression trees are for dependent variables that take continuous or ordered discrete values, with prediction error typically measured by the squared difference between the observed and predicted values' (2011, p14).

As a result, this allows the Artificial Intelligence machine to sort through information at pace and to create pathways and links between different forms of information as shown below in Figure 2. Data can thus be sorted and interpreted and fed back to the user in a way that is useful.

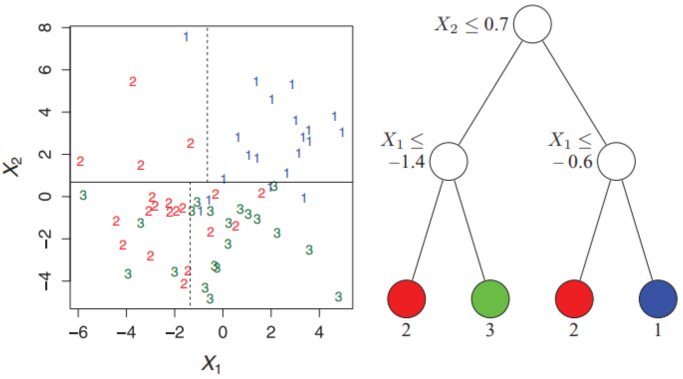


Figure 2: Random Forest [9]

Figure 2 shows us an example of a possible model that could be used in an algorithmic decision tree. Such a decision-making tree would be possible for some algorithms such as with a game of chess. There are a limited number of options and a limited number of moves. The system uses these options to come to the optimal outcome. This forms a base layer to the Artificial Intelligence system - to assist in directing training and decision making. Likewise, other forms of algorithms would be useful for different systems and for solving different problems. There are many complex forms of algorithms that serve different purposes (to explore all of these is beyond the scope of this paper). However, choosing the right algorithm for the problem that you are faced with can be very important - especially when it comes to the topic of explicability and how you enable transparent decision making.

It is important when considering Artificial Intelligence algorithms to consider the difference between an algorithm and a model. The term model in machine learning is ‘the output of a machine learning algorithm run on data. It represents what has been learned from “learning” the algorithm on the data and contains a specific set of functionalities of the algorithm’ (Haidar, 2021, NP). This model can then be used for future predictions/decisions on new data which can later be deployed and used in production.

In essence a machine learning model is data plus an algorithm or as Haidar puts it ‘Machine Learning Model == Model Data + Prediction Algorithm’ (2021, NP). As a result, when we consider ethics and AI it is important to also ensure that we consider, the algorithm, the data processes, any training setting that has taken place but above all the ethics of the data used.

2.1.2 Machine Learning

Machine learning is in many ways what it says on the tin. In essence machines can learn and can be ‘trained to make accurate predictions’ (Taulli, 2019,p 42). Bertolini, et al, define machine learning as ‘a branch of artificial intelligence that studies algorithms able to learn autonomously, directly from input data’ (2021, p1). At its most simplified machine learning is basically what you need to do to train ‘computers to “learn” patterns without being explicitly programmed for those patterns’ (Schuurman, 2019, p2).

[9] (Loh, 2011, p15).

Machine learning is used in a variety of fields such as recruitment, 'customer experience… customer service' (Taulli, 2019, pp47-48) and even in things such as Tinder and other dating apps (Taulli, 2019, p48). Machines are trained using data sets through what is known as supervised and unsupervised learning algorithms to fulfil a specific task.

There are many different forms of machine learning and a variety of models that are used. However, at their heart all forms of machine learning are 'capable of extracting knowledge from data, and continuously improve their capabilities, by learning from experience' (Bertolini, et al, 2021, p2). As we will see frequently in discussions around Artificial Intelligence - data is at the heart of much of what AI is about and this is also true when it comes to the subsection of machine learning.

It is important to realise that, as Tom Taulli succinctly points out, 'the goal of the machine learning process is to create a model, which is based on one or more algorithms. We develop this by training it. The goal is that the model should provide a high degree of accuracy' (2019, p49). In essence what you do in developing machine learning is to teach the model that when presented with X the answer is Y. And therefore, when presented with similar issues the machine can respond accurately.

Training a model is a key component to the development of machine learning. In our next section we will explore how this is to be accomplished.

2.1.3 Supervised Learning

Very often the first part of the process, in developing Artificial Intelligence (specifically machine learning etc), is to train the machine. The machine is taught with training data. This data is labelled by the author/developer so that the machine can develop a form of understanding or familiarity with the types of patterns that it is going to need to be able to recognise and this type is known as supervised learning.

When neural networks are used; data is fed into the machine (as seen in figure 2.2) after being labelled, cleaned, and organised and the neural networks see the myriad of connections and relationships between this data. Within the hidden layers this information is assessed and refined to produce an output. This process is refined as the machine is trained to come up with whatever the 'correct' answers could be in a given situation. For example, in chess you would correct the machine to ensure that it moved the chess pieces in ways that were legal according to the rules of the game.

For example, when using the supervised learning method, the machine learns the ‘relationship between the images and its labels based on the features’ (Hameed, 2019, p124). Even things such as numbers are broken down into their different components to identify an image by different numerical factors. These numerically based “images” are then used to identify different patterns that can be identified in a range of different data sets from numbers to speech, images of people’s skin to sifting through thousands of words in a document.

In the training stage the machine is given data (e.g. a set of images) and a suitable algorithm to run it with various training settings. By doing this the author can see how often the machine gets things right or not and is then able to tweak settings to ensure that the machine is able to do the job that it has been designed to do before being put into use (i.e., production). This training stage is key to the development of AI software.

If, for example, you wanted to develop a machine that recognises an image of a horse you would have to use labelled features to train the Artificial Intelligence what it is looking for. For example, patterns in the data, similarities, and what distinguishes this image from other images. The machine would also need to be able to recognise when something is not a horse. This is why we must train Artificial Intelligence; to develop accuracy.

There are many difficulties that can be encountered as a result of the training stage. For example, if the data is flawed or not labelled correctly then there can be potential for things to go wrong. Likewise, you could give the machine too much or too little training data and this would affect the performance of the machine once it is deployed. The training stage can be a serious downside to AI since the machine ‘may require hundreds of thousands or millions of hand-labelled examples’ (Taulli, 2020, NP). This is both time consuming and could also be a cause of potential issues due to human error. This signals that this training area of development should be of particular concern for ethicists and moral theologians as it is an area of great importance in determining how an artificial intelligence acts or behaves. Essentially the quality of the training creates the quality of the product both in terms of technology and morality. If the training data is biased the machine too will be biased, and it is issues such as this which lead to some of the greatest risks. There are not only issues with data and its bias but also training, as it is a very expensive process which requires supercomputers with specialised processing cores to run several rounds of training such as in the case of processing images and visual data. This has resulted in heavy computational burden which may affect the accuracy of models.

2.1.4 Unsupervised Learning

It is also possible to train a machine with unlabelled data (Schuurman, 2019, p2), but this requires lots more data which also needs to be cleaned and refined to ensure that it is suitable and free of outliers that may lead to poor outputs. In this form of training, you would first need to 'start with a huge amount of unstructured data… Then you will use a sophisticated model that will process this information and try to determine underlying patterns, which are often not detectable by people.' (Taulli, 2020, NP).

There are a variety of different methods used to produce unsupervised learning but 'the most common approach… is clustering, which takes unlabelled data and uses algorithms to put similar items into groups' (Taulli, 2019, p52). The enormous amount of data needed for effective and accurate Artificial Intelligence means that unsupervised learning will 'likely be critical for the next level of achievements' (Taulli, 2019, p53) in the field of AI. It makes life easier as the developer does not need to label each and every piece of data individually which saves a huge amount of time. As unsupervised learning develops and continues to 'detect and… extract patterns in… data' (Bertolini et al, 2021, p3) it can pull out details that could be missed by a human developer. While unsupervised learning could be useful to spot patterns that would otherwise be missed- there is also a higher risk of inaccuracy and mistakes in data and results/outputs.

2.1.5 Semi-supervised Learning

When you have some data that is labelled and some data that is not labelled it is possible to 'use deep learning systems to translate the unsupervised data into supervised learning' (Taulli, 2019, p54). In essence you can use Artificial Intelligence to clean the data that you are going to use for the production and training of other forms of Artificial Intelligence. Given the large amount of data needed to train accurate AI, it is unsurprising that there is a need to implement these processes to clean the data and order it as it would otherwise take a human person; days, weeks, and even months to do the same task.

2.1.6 Neural Networks

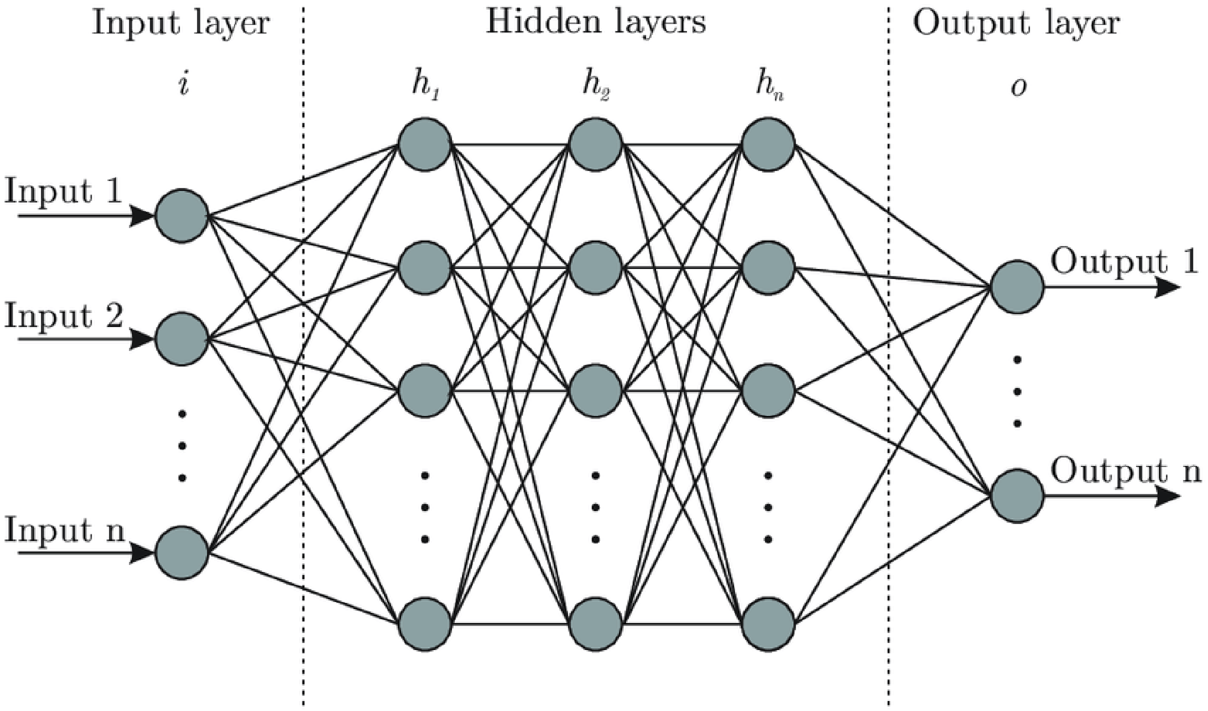


Figure 3: Neural Network [10]

Increasingly ‘neural networks have become a mainstay of artificial intelligence and cognitive science’ (Sun, 2014, p108). This form of programming is based on belief that ‘cognition emerges through the interactions of many simple processing elements or units (i.e. “neurons”)’ (Sun, 2014, p209) and connections between theses neurons that has various weights and thresholds. Figure 2.2 demonstrates how neural network work.

It is intended to work in much the same way that we perceive the human brain to work. Data is input on one side; it is run through various and competing calculations, and information points, and it is allowed to come to conclusions based upon the information that has been put into the machine (i.e., mainly some mathematical algorithms and parameters to tune the result) and the data that it has received.

[10] Shukla, 2019, NP

2.1.7 Deep Learning and Black Boxes

Deep learning is a specific type of neural network (Haenlein and Kaplan, 2019, p8). This form of neural network uses 'many layers of perceptions which can be trained using special techniques' (Schuurman, 2019, p2).

It is easy to confuse the topics of machine learning and deep learning. Deep learning is a 'subfield of machine learning' (Taulli, 2019, p71) and has more layers of neural networks and goes deeper than machine learning- hence the name. Yet it is important to remember that while machine learning is becoming more and more prevalent- deep learning is still at the cutting edge of AI development. In deep learning the developer will use neural networks and different layers of machine learning to mimic the way that we believe that the human brain works and therefore can be used for more complex analysis than machine learning can, for instance in the case of healthcare.

However, this leads to multiple different levels of hidden layers leading to the models in question being more difficult to comprehend or to explain how or why a model has come to the conclusion that it has. The issue with deep learning is that it is 'essentially a "black box". This means it can be nearly impossible to understand how the model really works' (Taulli, 2020, NP).

The benefit of deep learning is that 'Deep Learning methodologies allow working on almost raw data with little or no need for data pre-processing' (Bertolini et al, 2021, p21-22). Yet the fact that 'Deep Learning techniques and, more, in general, all the NN [Neural Network] based approaches, are difficult to be interoperated and could be negatively seen as a black box' (Bertolini, 2021, p22) remains an issue that leads to any use of such forms of AI being considered with incredible care. These concepts will be explored in more depth later.

2.1.8 Explicability / Interpretability

As we will see in later sections, the explicability and interpretability of models used in Artificial Intelligence development and implementation is a key area of ethical discussion. Transparency (another term that could be argued is interchangeable with explicability and interpretability) is often discussed in the codes of ethics employed and recommended by governments and other institutions when it comes to good practice in AI. Yet, as we will see, this is less straightforward than it may seem at first.

It is true that ‘many machine learning (ML) algorithms used to develop these systems are inscrutable, particularly deep learning neural network approaches which have emerged to be a very popular class of ML algorithm’ (Rai, 2019, p137). The hidden layers and multiple different unexplainable characteristics of these models can often lead to the machine being difficult to interpret or explain which can lead to a wide variety of ethical issues, as we will see later.

Explainable Artificial Intelligence is growing in frequency and popularity. As people more often ask why a machine has recommended something, so the demand for explanations grows. The useful category of explainable AI allows for users to gain ‘visibility into how an AI system makes decisions and predictions and executes its actions’ (Rai, 2019, p137-138). This would be particularly useful in areas such as social media, or retail (such as Amazon) as this allows users to make more informed and responsible decisions. For example, when Amazon says that because you bought X product, we recommend Y product based on people that have bought X have often also bought Y. However, the question remains as to whether this is the only information about a consumer that companies like Amazon hold and how much influence other factors have on AI decision making. The more information we provide to people about how AI works, the more ethical it becomes as individuals are then enabled to make more responsible decisions with their data.

The Turing Institute defines transparent AI as ‘the ability to know how and why a model performed the way it did in a specific context and therefore to understand the rationale behind its decision or behaviour… transparent AI involves the justifiability both of the processes that go into its design and implementation and of its outcomes’ (Leislie, 2019, p35). In essence, this advice recommends that every step of AI design and development be transparent and explainable.

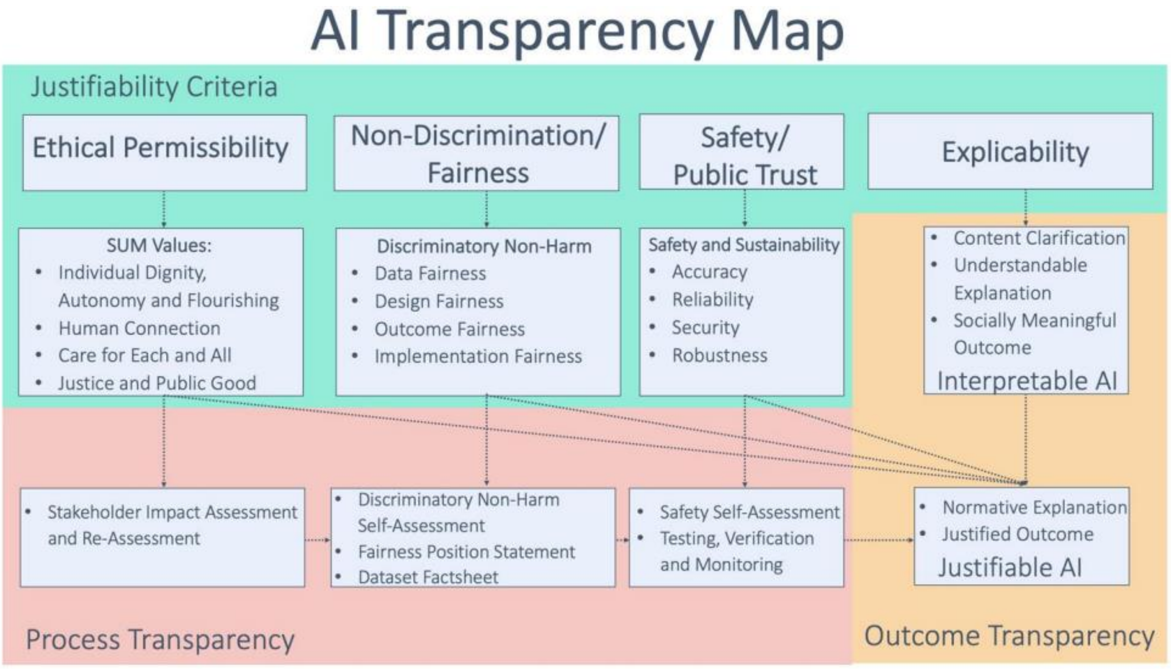


Figure 4: AI Transparency Map [11]

As in figure 4, by the Turing Institute, points out - there are many layers to transparency and explicability. There is more to be considered than simply explaining how and why the model comes to the conclusions that it does. As we will see when discussing current ethical codes and legislation - there are wider factors that must be considered around fairness and avoidance of discrimination. Designers and developers of AI do need to be concerned with accuracy and vetting but, as we shall make clear in later arguments, they must also recognise that they are responsible for that which they create. To be transparent and explainable the motives and decisions must be justifiable and responsible at every stage of the process by everyone involved.

2.1.9 General Artificial Intelligence

As we have seen, many of the concerns among theologians, philosophers, and others tasked with the ethics of AI, is the concept of General Artificial Intelligence.

General Artificial Intelligence, or otherwise known as Artificial General Intelligence, is often referred to as strong AI. This form of Artificial Intelligence ‘refers to the effort to create machines that are able to tackle any problem by applying their skills. Just like humans they can

[11] Leislie, 2019, p36

examine a situation and make best use of the resources at hand to achieve their objectives' (Ashri, 2020, p17).

In our considerations of Artificial Intelligence in terms of Neural Networks, Machine Learning, and Deep learning; while AI can do many wonderful things, it is limited by the projects to which it is committed and designed for. The idea of General AI is more closely linked to the depiction of AI that is often seen in movies or TV shows, such as *Person of Interest*, or *I Robot*, where we see an AI that can think freely and learn to do more general tasks. While this is not necessarily impossible and could well be an issue that requires ethical discernment in the near future; it is an issue people are aware of and studying but often at the expense of the ethical issues presented by AI as it already is here and now.

While this study does recognise that there is a clear and discernible need for ethical reflection on this topic that it is

1) Beyond the scope of this project, and

2) That there is a lack of ethical consideration, from a Christian perspective, on the ethical consequences of current and already existing AI products, services, and other uses.

AI is already with us and if we only concentrate on the future potential issues, we will miss the issues and complexities that we are already being faced with.

2.1.10 Ways of defining Artificial Intelligence

Artificial Intelligence is currently used in a variety of fashions: for example, it has been used to great effect in chess (by beating the world champion), in mathematics, in games and in many other areas (Franklin, 2014, p23-24). This is particularly true in medicine and healthcare where AI has the potential to provide real practical help in the diagnosis and treatment of a wide variety of issues. This is an area that we will explore in more detail later.

Today, Artificial Intelligence is also particularly prevalent in the area of 'data mining' (Franklin, 2014, p28) which allows people to sort through vast quantities of information quickly and efficiently and to increasingly move towards a form of 'cognitive computing' (Franklin, 2014, p29) where the machine could not only say what it was doing but 'why it was doing it' (Franklin, 2014, p29).

However, there are clear limitations to what is currently possible with Artificial Intelligence. Artificial Intelligence, as it currently stands, ‘falls short of human capacities in some critical sense’ (Bostrom and Yudowsky, 2014, p318). The human quality that many see as lacking is ‘generality’ (Bostrom and Yudowsky, 2014, p318). AI can be used in a variety of ways today; however, each machine is mostly designed and used in a very specific way for a very specific purpose. The machine is designed for that specific task or that specific activity. For example, the machine that beat a world champion chess player- Deep Blue- is brilliant at chess: but this same machine would not be able to tell you how to do something as simple as use a toaster. As we have seen, many ethicists and moral theologians are very often concerned with the threat of what Artificial Intelligence could become or the results of the development of ‘general AI’. However, for the time being at least, this appears out of reach and thus we should focus on practical and current uses of Artificial Intelligence technology. This is not to diminish the potential threat that many are concerned about but to realise that AI, as it currently stands, is already changing the way we live and work and exist. For example, researchers from the University of Oxford predicted that ‘47 percent of US jobs are at risk of being replaced by Artificial Intelligence technologies and computerization’ (Schuurman, 2019, p3). This is the real world and current ethical dilemma faced by ethicists, moral theologians, and computer scientists that needs to be considered before we think about the potential risks of developments that have not even occurred yet.

2.2 Areas in which Artificial Intelligence is currently used.

Artificial Intelligence is used in just about every sphere of modern life. Below we will explore some of the areas in which it is currently in use. This is not an exhaustive list by any stretch but is offered to explore the key areas in which AI is used so that we can identify common ethical issues that need to be considered in the development of a responsible ethic of Artificial Intelligence development and implementation. This is particularly important when we consider that Artificial Intelligence is now influencing ‘what we buy, who we hire, who our friends are, what newsfeed we receive, and even how our children and elderly are cared for’ (Rai, 2019, p137). Likewise, when we consider the areas in which AI is currently used, it becomes clear how important a functioning and responsible ethical approach is to the development and implementation of this new technology.

2.2.1 Healthcare

When we consider the use of Artificial Intelligence in healthcare the most obvious thing that will come to mind for many is the concept of a virtual doctor or an AI machine diagnosing someone. The concept of AI performing diagnoses and assisting with a diagnosis is not impossible and there are examples of this happening. However, Artificial Intelligence can often be more useful in the more mundane things within a healthcare setting. For example, sorting and organising medical records, ensuring that data is safe, and spotting patterns across vast datasets to notice what individual human persons would fail to do.

Even in the more mundane areas of healthcare AI is a useful tool. In *Artificial Intelligence Simplified*, George and Carmichael look at how AI can be used in something like scheduling for surgeries to ensure that things arrive in the correct place at the correct time to avoid issues, delays, and mistakes (2021, pp22-27). These things might seem trivial but having the correct anaesthetic in the right theatre with the doctor that is supposed to be there will save lives, make things more efficient, and decrease the chances of human error.

However, there are developments that suggest that Artificial Intelligence could be used even in the diagnosis, prevention, and treatment of a variety of conditions. This could be particularly helpful when looking at conditions such as cancer where early detection is critical. AI systems are often good at spotting things that humans could well easily miss. Likewise, other common factors or behaviours, of those who later go on to develop cancer, could be identified by AI machines that would not necessarily be spotted by individual doctors.

However, we must be careful - especially when we consider how sensitive the whole area of medicine is. We have seen how important getting data right is to Artificial Intelligence - and likewise, we will see that the risk of bias is an ethical risk that we should not downplay and could have catastrophic impacts if used without proper caution and due diligence.

As in many areas in which we use Artificial Intelligence - the machine is a tool to be used to improve the industry and the lives of individuals. It should not be envisioned, when used responsibly, that such technology is intended to replace the essential work of medical professionals but rather to assist them in spotting things earlier and improving the quality of treatment that they can provide.

2.2.2 Agriculture

Agriculture and farming are not areas that one would usually associate with Artificial Intelligence technology. Yet, the prevalence of AI in these areas demonstrates clearly how these emerging technologies are impacting every area of human life.

Artificial Intelligence technology has the potential to assist in our efforts to reduce and even eliminate hunger and starvation. AI can be used for calculating the best crop rotations or which chemical additions should be added to the soil to produce more and higher quality produce both in indoor and outdoor settings. Artificial Intelligence will consider factors that the human mind may not even consider, and this could be of great benefit in many parts of the world.

One striking example of the use of AI in agriculture are the ‘RoboBees’ (Wyss Institute, 2022, NP) that are currently being developed in the USA. These tiny robots are designed to work together to fulfil various tasks such as ‘environmental monitoring’ and ‘crop pollination’ (Wyss Institute, 2022, NP). As the number of bees decreases in the world- this could be one way to ensure essential pollination occurs in areas that could struggle because of this. Likewise, there is potential for such robots to be able to monitor large areas of crops on even a minute level - saving time and ensuring less issues with produce.

While RoboBees might not be in widescale use already there are areas in which Artificial Intelligence is already used. AI programmes are used to ‘monitor crop moisture, soil composition, and temperature in growing areas’ (Young, 2020, NP). At the same time, while AI can be used to mitigate the consequences of global warming on agriculture it is also true that, as Sydney Young points out, Artificial Intelligence development has the potential to cause climate change to get worse. For example, ‘due to the large amount of data that AI needs to process, training a single AI releases five times the emissions that an average car would emit during its lifetime, thereby adding to the already substantial environmental impact of computing technology’ (Young, 2020, NP).

Artificial Intelligence can be a great asset in out attempts to assist with world hunger and to mitigate the damage done to the environment due to climate change. However, developers, researchers, and corporations must also take seriously the environmental impact of such technology due to the large carbon emissions caused by data storage and machine training as we will explore later.

2.2.3 Social Media

It is nearly impossible, in one way or another, in the modern world, not to be engaged in some form of social media. What people may not realise is that today 'Social media platforms increasingly use powerful artificial intelligence (AI) that are fed by the vast flows of digital content that may be used to analyze [sic] user behavior [sic], mental state, and physical context' (Lewis and Moorkens, 2020, p1). We have allowed social media companies free range with vast amounts of our own personal data, and this is than used to advertise to us, and to fuel ever better AI prediction models.

It is often said by people that they think of something they might need or want and then for the next few days there are adverts for said product everywhere that they access online. This is not a random coincidence. Artificial Intelligence technology is used extremely effectively to spot trends and to tailor them into advertising. At times this might be helpful and clearly understandable. For example, a website might say you have bought X product and therefore it suggests Y product because this is a product that people that have bought X often also buy. However, such tactics are not always explained or made clear. This is especially risky when Artificial Intelligence can be responsible for assisting in the radicalisation of individuals through the creation of echo chambers. AI shows you what it thinks you want to see and therefore, when used in the context of social media, it can make it seem like everyone shares your own views or perspectives. This could have serious risks by helping to cause someone to become politically, religiously, or ideologically radicalised because all the individual is being presented with is a single narrative.

There is, at present, little regarding 'accountability, data governance, design-for-all, human oversight, respect for human autonomy, and transparency' (Lewis and Moorkens, 2020, p1) in the current guidelines that are enforceable against social media companies, and this is a large risk when it comes to society. One only has to look at the reports that Facebook allegedly sold 87 million people's data during a presidential election to Cambridge Analytica (Criddle, 2020, NP) to see the risks.

Social media has connected us like never before to one another, but it also runs the risk of separating us more than ever and dividing us into our own little groups. Artificial Intelligence needs to be used responsibly in this space to ensure that social discourse is not reduced to what we read on social media.

2.2.4 Automated Cars

Self-driving cars might sound like something that is incredibly futuristic. However, they may not be as far from reality as might once have been assumed. Cars with limited capacities for self-driving already exist, for example Tesla autopilot (Rimell, 2022, NP). The government of the United Kingdom has already looked at the possibility of allowing 'self-driving vehicles to be used for public transport or deliveries … from 2025 – and wouldn't require anyone on board to possess a full driving licence' (Rimell, 2022, NP).

As much as self-driving cars present a wonderful image - they can also be seen as being a massive ethical issue in themselves. For example, if there is a crash who should be held responsible. Even the classical question of the trolly theory comes into play when it comes to automated vehicles. In an accident where someone will die no matter what the Artificial Intelligence machine chooses to do- who should it protect and who should it allow to die and how does it justify that decision. It is clear, more than anything, that the biggest ethical risk when it comes self-driving cars, and many other forms of Artificial Intelligence, is the question of public trust (Smith, 2020, p682).

There are many areas in which Artificial Intelligence can be used to augment and assist our transportation needs - especially in the delivery of goods and materials. However, this is also not a straightforward ethical question. The ways in which it is programmed to make what are in essence ethical questions is essential. As we have often seen - the question of responsibility again raises its head. Who is to be held responsible if there is a crash involving an automated vehicle if it appears that the crash is the fault of the Artificial Intelligence?

2.2.5 Education

When we consider the role of Artificial Intelligence in the world today and in the future, we rarely think about education. However, there are many ways in which education is affected by the increased use of AI in the modern world. One way in which this is most striking is the paradigm shift needed to prepare students to enter a very different workforce than the one that their parents and grandparents grew up in. Universities that see their job as being to 'prepare students for fulfilling – and successful – roles in the professional world' (Aoun, 2017, pXV) will struggle as Artificial Intelligence changes the ways that we work and interact with the world around us.

More immediately, during the course of writing this thesis, *ChatGPT* has been launched by Open AI. Open AI says that they have:

‘Trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer follow-up questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests’ (OpenAI, 2023, NP)[12].

This is a big step, and all the consequences of this development cannot be discussed in this thesis due to a lack of time. This is particularly important because this Artificial Intelligence Machine seems to be able to do some things, such as essay writing, which could not have been done as easily before. However, it has already been noted by academic professionals that this machine has been used already by some students in assignments. There are ways of detecting the presence of AI written papers, however, it does demonstrate a real risk to the education sector. Artificial Intelligences such as ChatGPT could be very useful in the field of education. However, AI does pose many questions that the education sector will have to face – from preparing students for the workforce to how assessments are conducted.

2.3 Conclusion

In this chapter we have seen some of the history of Artificial Intelligence technology - how it has been developed and some of the ways in which it works. There is far more to how Artificial Intelligence machines are designed and a large variety of other ways in which modern AI technology can be created and implemented. This chapter is far from exhaustive in this regard and should not be regarded as such. Likewise, we have observed how the future of AI could progress and we have begun to look at some of the ethical issues that present developments create. Finally, we have looked at a very small number of the ways in which AI is used in the modern world.

Artificial Intelligence is all around us. It is used in everything including, cars, smart phones, social media, and even in the processing of food. There are so many ways that this technology is being utilised and implemented to improve our lives. Yet, hopefully this chapter has also shown how the ethics of AI is more complicated, and more pressing, than many of the ethical questions of the past. This is particularly true when it comes to explicability and responsibility when it comes to AI. In the past individuals were clearly responsible for their

[12] As on the website

own actions. Ethics could, in many cases, be seen as intention followed by act followed by consequence. As with many things in the world today this is no longer a straight line and is complicated by globalisation. As we will see AI, in particular, does not fit these classical ethical frameworks and therefore it needs to be considered who is responsible and how can we be ethically responsible when it comes to the development and implementation of Artificial Intelligence Technology.

Chapter 3: Current Codes of Ethics in use today

There are a range of differing perspectives and attitudes towards Artificial Intelligence ethics today. These can often be complex and differ significantly from organisation to organisation. There is also a ‘relationship between ethics and law’ (Boddington, 2017, p25) that needs to be taken into consideration. The complexity of this consideration is added to when you consider that organisations involved in the development and implementation of Artificial Intelligence technology also need to consider the fact that ‘there may be great differences in some aspects of the law between different jurisdictions’ (Boddington, 2017, p25). The way in which an ethical framework should interact with law on both a national and international level is beyond the scope of this work; however, legal frameworks are extremely significant in any study of responsibility and Artificial Intelligence. For example, an ethically responsible government takes responsibility for AI and engages in the creation of legislation to protect citizens while still allowing for technological development.

The guidelines produced by different national and international governments and organisations are significant in demonstrating where the current frameworks are in terms of legislation and guidance. Such frameworks can help in identifying where the concept of Responsibility could be beneficial to the ongoing discussion of Artificial Intelligence ethics. A selective and representative framework has been used in this thesis due to time and size. As a result, it will not be possible to give an in-depth assessment of all the different ethics policies and principles and therefore this thesis has focused on a small number of such regulations/recommendations to pull out common themes.

At a governmental level, both nationally and internationally there is a need to develop codes of ethics for Artificial Intelligence. Many codes have been created, some of which, will be outlined below. Often these codes are in some way related to the law, but they also often include suggestions and recommendations around best practice rather than strict legal frameworks which, as we will see, risks the consequences of ineffectual AI ethics.

3.1.1 United Kingdom

The UK Government has a page dedicated to ‘AI ethics and safety’ (UK Gov, 2019, NP). On this page the Government recognises that ‘The field of AI ethics emerged from the need to address the individual and societal harms AI systems might cause’ (UK Gov, 2019, NP). The government of the UK also identifies three areas that it deems to be the biggest areas of risk regarding AI development and implementation. These are:

1. ‘Misuse - systems are used for purposes other than those for which they were designed and intended.
2. ‘Questionable design - creators have not thoroughly considered technical issues related to algorithmic bias and safety risks’.
3. ‘Unintended negative consequences - creators have not thoroughly considered the potential negative impacts their systems may have on the individuals and communities they affect’ (UK Gov, 2019, NP).

These risk factors are interesting in that they focus more on unintended consequences than societal impacts or questions of human dignity/common good that could arise from AI development and implementation. However, later the page does talk about the need for AI to be ‘fair and non-discriminatory’, ‘worthy of public trust’ and to ‘prioritise… transparency’ (UK Gov, 2019, NP).

The UK Government website also points especially to a document from the Alan Turing Institute: *Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector* (Leislie, 2019). Within this document, David Leislie discusses the ‘SUM Values’ (2019, p9). These are:

1. ‘Respect- the dignity of individual persons’.
2. ‘Connect- with each other sincerely, openly, and inclusively’.
3. ‘Protect- the priorities of social values, justice and the public interest’.
4. ‘Care- for the wellbeing of each and all’ (Leislie, 2019, p9).

Unsurprisingly, the webpage and this document from the Alan Turing Institute both focus heavily on ‘public trust’ (Leislie, 2019, p5) (UK Gov, 2019, NP). Accountability and transparency are also heavy features of these sources. Much of this comes down, as we have seen repeatedly, to data and ‘data fairness’ (Leislie, 2019, p14). The document from the Alan Turing Institute, along with many other sources, makes it clear that ‘responsible data acquisition, handling, and management is a necessary component of algorithmic fairness’ (Leislie, 2019, p14). Simply put, bad data equals bad AI. As we have already seen, data is an essential element of AI and must be a key ethical consideration going forward. Likewise, even here in this document there is a clear awareness that we must take ethical responsibility seriously.

It is however notable that two significant areas are not addressed within the webpage or the document. Firstly, there is no direct legislation relating to these issues when it comes to Artificial Intelligence. These are by and large recommendations and suggestions. There are elements that are covered by other pre-existing legal principles such as human rights and data protection legislation. However, this is unlikely to consider the complexity and future repercussions of AI. Secondly, the impact of AI on jobs and employment is not addressed in any detail if at all. There is consideration within the Alan Turing Institute document regarding fairness and (Leislie, 2019, p16-18) ensuring that persons can ‘make free and informed decisions about their own lives’ (Leislie, 2019, p10), however, this does not address the significant issue that could be presented in the workforce because of AI. There is an overwhelming sense that AI is a positive force and that those who are creating AI are doing so for positive motives within the document and webpage. Sadly, this is not always the case and there should always be a concern that a pursuit of profit/profitability does not outweigh our considerations regarding the dignity of the human person and their responsibility such as the recent case where Microsoft reduced the number of employees committed to working towards AI ethics, safety, and responsibility (Sankaran, 2023, NP).

Likewise, the concept of responsibility can seem unclear in the documents. It is imperative that legislation be paired with responsibility so that responsibility for AI becomes something that we all do together and not something that is left in the hands of a small number of people.

3.1.2 European Union

The EU guidelines on Artificial Intelligence ethics states that 'trustworthy AI should be:

(1) lawful - respecting all applicable laws and regulations

(2) ethical - respecting ethical principles and values

(3) robust - both from a technical perspective while taking into account its social environment' (EU, 2021, NP).

Furthermore, the EU stipulates that, for AI to be trustworthy, AI developers should meet '7 key requirements' (EU, 2021, NP). These are:

- 'Human agency and oversight' (EU, 2021, NP).
- 'Technical Robustness and safety: AI systems need to be resilient and secure' (EU, 2021, NP).
- 'Privacy and data governance' (EU, 2021, NP).
- 'Transparency' (EU, 2021, NP).
- 'Diversity, non-discrimination and fairness: Unfair bias must be avoided' (EU, 2021, NP).
- 'Societal and environmental well-being: AI systems should benefit all human beings, including future generations' (EU, 2021, NP).
- 'Accountability: Mechanisms should be put in place to ensure responsibility and accountability for AI systems and their outcomes' (EU, 2021, NP).

These guidelines are designed for self-assessment and rely heavily on the good will of governments, companies, and other organisations for their application. There is a clear overlap with legal requirements in some areas. For example, privacy law such as GDPR regulations, laws against discrimination and other business laws are regulations that play a part in the lives of different organisations. However, although these guidelines are noble and important, they do little to ensure compliance and is also more reactionary rather than looking at the core ethical principles that are at stake. Why should companies and individuals follow these recommendations? And what is the framework in which these ideas fit? There is no legal basis for responsibility within these guidelines which in many ways makes such governmental systems toothless. Likewise, responsibility is highlighted as being important and yet nothing is done to identify who is responsible and what responsibility looks like in practical terms. Ironically it is the lack of deontological framing that makes responsibility impossible under these circumstances. To be responsible, in the sense that this paper is arguing, one must first be able to internalise responsibility which is arguably impossible if there is no external clear deontological law.

3.1.3 United States of America

The American *AI initiative* that was launched by President Donald Trump came up with 10 principles to be used in developing a strategy within the USA (Hao, 2020, NP). These principles have been released by 'the White House Office of Science and Technology Policy (OSTP)' (Hao, 2020, NP) and were designed with three main goals in mind:

1) 'To ensure public engagement'.

2) 'To limit regulatory overreach'.

3) To 'promote trustworthy AI that is fair, transparent, and safe'.

In this case there is more of a move towards the concept of actual practical regulation and consequences but in the end returns to principles as listed below.

The principles found within the document are as follows:

1. **'Public trust in AI.** The government must promote reliable, robust, and trustworthy AI applications.

2. **Public participation.** The public should have a chance to provide feedback in all stages of the rule-making process.

3. **Scientific integrity and information quality.** Policy decisions should be based on science.

4. **Risk assessment and management.** Agencies should decide which risks are and aren't acceptable.

5. **Benefits and costs.** Agencies should weigh the societal impacts of all proposed regulations.

6. **Flexibility.** Any approach should be able to adapt to rapid changes and updates to AI applications.

7. **Fairness and non-discrimination.** Agencies should make sure AI systems don't discriminate illegally.

8. **Disclosure and transparency.** The public will trust AI only if it knows when and how it is being used.

9. **Safety and security.** Agencies should keep all data used by AI systems safe and secure.

10. **Interagency coordination.** Agencies should talk to one another to be consistent and predictable in AI-related policies.' (Hao, 2020, NP).

It is notable that these principles suggest more practical steps than many of the others that we have considered such as 'risk assessments' (Hao, 2020, NP) and 'public participation' (Hao, 2020, NP). Yet these principles are clear that agencies and organisations are to make decisions for themselves. While these do provide a good framework, they are recommendations at best and a free pass for organisations to do as they see fit at worst. There is a clear concern against regulatory overreach within this system, and a concern for profit, yet it makes no attempt to create any form of meaningful legislation or accountability. The engagement of the public is imperative for the development of responsible AI as is public trust. However, unlike the EU document, this document does not make any reference to accountability or responsibility. Much of the concern is consequentialist rather than deontological here and that is a concerning trend when the consequences are not yet clearly know.

There are many other ethical systems in use today and each company and government have a slightly different approach. However, the above examples do highlight the common themes that we find in each. Yet, the issue in this is that many, if not all, use a principle-based, and not legislatively backed, approach as outlined above. As we will see next - this is, however, far from a straightforward or necessarily good/ useful practice in some ways.

As much as principles are necessary- it is far too easy for principles to simply remain principles and to have no real impact upon life.

3.2 Conclusions: The issue with principles-based approaches

In a paper, on the topic of Christianity and AI, from *The Trinity Conference 2022: A Catholic Response to the Digital Age*, it is argued that the issue with a principle-based approach to Artificial Intelligence ethics is threefold:

The first issue with principled based Artificial Intelligence ethics is that it is often left simply with developers and with companies. There is a great desire for companies to self-regulate to ensure that regulation does not stifle innovation. However, this is risky. There is always a 'temptation for businesses and institutions to put profit over the common good.' (Nelson, 2022 B, p52) and as a result there is a need for some form of regulation that goes beyond simply looking at principles.

The second issue that was discussed is that without oversight it is easy for deliberate or accidental bias to creep into any project. This is especially true when owners and managers of companies that use Artificial Intelligence technology may not understand the processes involved in AI development. This lack of oversight paired with principles-based ethics leaves developers alone responsible for what is created when this responsibility should be wider reaching. As Bonhoeffer argues – we are all in some way responsible and need to take responsibility seriously as our vocation.

The final issue that was discussed is the complex task of how principles would be 'translated into practice' (Nelson, 2022 B, p53). Principles that we come up with still must be used and put into practise. When we leave things as simply high-level principles, without any means of putting them into practice we do not give people the tools that they need to ensure ethical Artificial Intelligence[13].

Without principles as a starting point, it is difficult to get any ethical consideration off the ground. However, there is a clear danger that such principles remain simply as principles

[13] These paragraphs relate to the paper:
Nelson, J. (2022 B). The Rome Call to Artificial Intelligence Ethics Inside the mind of the Machine: How the Church can respond to the ethical challenges presented by AI. *The Trinity Conference 2022: A Catholic Response to the Digital Age.* Edited by Hayward, H, and Stacey, G. Leeds Trinity University, UK.

and do not have any impact on the behaviour or practices of organisations that use Artificial Intelligence technology.

To be responsible, as we have seen, means to act. To be responsible means to engage with the world and not to allow our ethics to remain without substance and practice. An ethics of responsibility will not leave principles as they are or even accept them on face value but seek to apply them in a way that is appropriate and relevant to the situation in which one finds oneself.

Many such principle-based approaches have created ‘vague, high-level principles and value statements which promise to be action-guiding, but in practice provide few specific recommendations and fail to address fundamental normative and political tensions embedded in key concepts’ (Mittelstadt, 2019, p1). The problem, above all, with many of these governmental and international codes of ethics is the fact that they are often best practice or recommendations - when there a need for further action such as international treaties or laws. Law must be added to responsibility in order for it to be effective.

To be responsible, it is, at times, the obligation of the state to step in and legislate with such new technology to curb abuses and ensure safety.

In Artificial Intelligence ethics: ‘the truly difficult part of ethics - actually translating normative theories, concepts and values into ‘good’ practices AI practitioners can adopt is kicked down the road like a proverbial can’ (Mittelstadt, 2019, p6). On the one hand one could consider that such a critique also applies to our concept of responsibility. However, while it is true on one level that responsibility alone is not sufficient, it is also important to recognise that a key part of an ethics of responsibility must be to call upon legislators to be responsible and to act to hold organisations and individuals accountable for that to which they are responsible for in the development and implementation of Artificial Intelligence. A responsible government cannot avoid its responsibility to act to protect the privacy, safety, and security of citizens and their data.

3.3 How Bonhoeffer’s conception challenges current ethical frameworks

Current ethical frameworks, as we have seen above are set upon the basis of principles. Many of these are envisioned to work much as the principles of medical ethics have traditionally worked. However, when a new field of study, such as computer science and the discipline of Artificial Intelligence, there is a clear need for a different way of approaching ethical questions. This is particularly complicated where areas of ethical concern overlap- as we have seen in the case of Medicine and Artificial Intelligence. The answers and solutions are far from straightforward, at least at the moment. This is therefore a time in which developers, business managers, governments, and individuals must take responsibility and stand together for the betterment of humanity.

Bonhoeffer's ethics of responsibility can help to move the discussion forward by emphasising the centrality of internalisation of ethics and ethical codes on the part of the developers and implementers of Artificial Intelligence technology, while on the other hand pushing for us all: governments, individuals, companies, and developers, to consider our role in technological developments and to act responsibly. In many ways it can be argued that governments have failed to take responsibility for Artificial Intelligence. Yes, they have produced guidelines and recommendations, but little of it is enforceable. This is further complicated by the fact that there are overlapping legal and moral concerns. However, such complexity requires a special level of responsibility to be taken as a serious concern on the part of individuals and organisations.

However, things are moving in the right direction. The UK government document *Establishing a pro-innovation approach to regulating AI An overview of the UK's emerging approach* does begin to look at how the government could begin to take responsibility and more importantly- define who is actually responsible by saying 'accountability for the outcomes produced by Artificial Intelligence and legal liability must always rest with an identified or identifiable legal person - whether corporate or natural' (Dorries, 2022, p14). However, more needs to be done to make such things clearer, more accountable, and ultimately to foster real responsibility.

Chapter 4: Issues and Applications of Bonhoeffer

4.1 Black Boxes and the challenge of Explicability Explored

Black boxes and explicability are key areas that Artificial Intelligence ethics will have to contend with and ensure an adequate solution is found for. Making AI that is explainable and transparent is becoming an ever more important quest yet at the same time this is less than straightforward.

A lack of transparency and explicability can lead to users rejecting the use of Artificial Intelligence, or worse, the use of unexplainable models and black boxes poses the risk of obfuscating 'the discovery of algorithmic biases arising from flawed generated processes that are prejudicial to certain groups. Such biases have led to large-scale discrimination based on race and gender in a number of domains ranging from hiring to promotions and advertising to criminal justice to healthcare' (Rai, 2019, p137). Such risks and instances run the risk of causing the rejection of Artificial Intelligence by government and individual consumers. Beyond this such things have clear ethical issues that must also be addressed. To be responsible, it is imperative that we recognise our biases (the reality of the situation) and that those involved in the development of AI ensure that such instances of discrimination are avoided at all costs.

This section will see how transparency and explicability can be achieved, why these are essential components to a Bonhoeffer based ethic of responsivity in Artificial Intelligence development and implementation and will consider ways forward to achieve explicable and transparent AI. Likewise, we must question whether it is responsible for black box models to be used at all as we will see next.

There has, lately, been an increasing trend 'in healthcare and criminal justice to leverage machine learning (ML) for high-stakes prediction applications that deeply impact human lives' (Rudin, 2019, p1) and may be used in the future in many other areas as we have already seen. As a result, there is a push for models, especially models used in such circumstances, to be made in such a way that they can be transparent, responsible, and explicable. However, creating such explainable and transparent models is far from straightforward. Yet, it is also clear that 'providing users with an effective explanation for the AI system's behaviour can enhance their trust in the system' (Rai, 2019, p137). Such risks can theoretically at times be justified. At the same time any such risk always must be weighed up against the obligation to be responsible and to act responsibly - even when it is not your individual responsibility per se (i.e Vicarious Responsible Action) as Bonhoeffer suggests.

Explicability can be achieved in a variety of ways. For example, in some cases it has become popular to create a model that explains how a given black box model works whilst still retaining the black box. As Rudin argues- this can be problematic and at times also unreliable (2019, p1). It is increasingly argued that we should do away with black box models and instead use models that are 'inherently interpretable' (Rudin, 2019, p1) or in another way to 'convert black-box models to glass- box models' (Rai, 2019, p138).

It is important that responsible developers consider if a black box model is even necessary for the task to which it is intended. For example, 'when considering problems that have structured data with meaningful features, there is often no significant difference in performance between more complex classifiers (deep neural networks, boosted decision trees, random forests) and much simpler classifiers (logical regression, decision lists) after pre-processing' (Rudin, 2019, p2). However, this is not a statement that is agreed by everyone. Some scholars, such as Rai, argue that deep learning does indeed 'sacrifice transparency and interpretability for prediction accuracy' (Rai, 2019, p138).

The risks presented by Artificial Intelligence again leads us back again to Bonhoeffer's warning against the 'violation and exploitation of nature' (Bonhoeffer, 1949/2009, p116). Artificial Intelligence risks changing very many things. Algorithms in use today- particularly in things such as social media can create echo chambers and restrict people's ability to engage critically and constructively with the world around them. Therefore, interpretability of AI models is so important- especially when it comes to anything that is engaged with human persons. Not only can things be unreliable- but they also have real world consequences. If human persons cannot understand how the AI system is working then it is even more difficult to avoid bias and to ensure safe, secure, reliable, responsible, and ethical AI.

When we consider responsibility within the context of black boxes it becomes abundantly clear that transparency and explicability are the only way forwards if this technology is to be used at all in a way that is responsible. If they are to be used, then the requirement for someone to take responsibility for the consequences of the machine becomes even more essential than before.

4.2 The issues with Data: Ethics Explored

Data, as Tom Taulli argues, is the 'fuel for AI' (2019, p19). As we will see, data 'is extremely powerful and critical for AI' (Taulli, 2019, p20), however, data is also the area of Artificial Intelligence development that carries the highest levels of ethical risk and the highest levels of challenges. This is especially true when we consider that in the development of AI 'the amount of money shelled out on data is enormous' (Taulli, 2019, p28). In our consideration of data, we will, by necessity, also be considering the topics of privacy, bias, safety, and security. Any attempt to consider an ethics of AI without looking at data would be impossible - as it is in data that we encounter our biggest dilemmas and challenges. Responsibility is particularly relevant here because it applies to the whole process:

- How data is sourced
- The responsible storing of data
- Responsible usage of data
- The prevention of bias
- The responsible structuring of data

And this is to name just a few. Therefore, data must be considered as being at heart of any attempt to find a responsible ethics for AI. 'All data has issues' (Taulli, 2019, p33) but it is important that responsible developers of AI, who see responsibility as a vocation, play a part in ensuring that the risks presented by data, its collection, and use, are minimised, and responded to appropriately.

As Gry Hasselbalch states

'A recent and new development is the transformation of all things into data as an effortless, costless and seamless extra layer of life and society. Data at the time of writing no longer just captures politics, the economy, culture and lives - data is their extension' (2021, p1).

Everything is being transformed into data in our modern world, in part, to fuel AI. Therefore, what this means is that the bedrock of any attempt to develop an ethics of AI must first and foremost be an ethics of responsibility for and with data. Thankfully, this is an area that is being developed but is also an area that needs to be focused upon in order to develop an ethics of AI.

Tom Taulli proposes that when looking at data we should ask the following questions:

- ‘Is the data complete? What might be missing?
- Where did the data come from?
- What were the collection points?
- Who touched the data and processed it?
- What have been the changes in the data?
- What are the quality issues?’ (2019, p32)

Responsible data sourcing and management are key to the safety and security of Artificial Intelligence ethics. If the sourcing and management of data is not done properly - nothing else will be able to be ethical. If the ethics of data is good and responsible - much of the battle for ethical AI is already won. For example, as we see in chapter 1, we must recognise the reality of the situation in which we find ourselves, such as our innate biases, we must see it as our vocation to ensure data is being stored and collected responsibly, and we must always consider how the decisions that we are in any way involved in will affect others.

As we have noted, to ensure responsibility, we must begin, as would Bonhoeffer, with acknowledging the reality of the situation in which we find ourselves today. Data is constantly being collected on individuals every moment of every day in the modern world. Our social media, supermarket loyalty cards, smart phones, and even in some cases cars are collecting data on us constantly; just to name a few. Companies will often use this data, or sell this data, for Artificial Intelligence training and development. As we have seen, good data management can make a world of difference to the reliability and effectiveness of AI. However, managing data responsibly, or irresponsibly, has many consequences into other areas. Most of these ethical issues come back to data and therefore responsible management of data is where many of the key ethical questions must be asked.

In this section we will look at the topics of privacy, bias, safety, jobs, climate, and how our ethical framework of responsibility can help to develop safe, secure, and responsible Artificial Intelligence.

4.3 Privacy

The concept of privacy is intertwined deeply with these questions. When dealing with data - especially data relating in any way to persons - the ethical consequences of privacy are a significant risk. It is clear that ‘Bonhoeffer supported the right to human privacy. Bonhoeffer stated this need for privacy walls in firm fashion... Bonhoeffer would not approve of the blurring of public/private information so common in contemporary culture’ (Arnett, 2005, p216). This is something that we, in modern society, can easily lack - for the significant reason that we rarely know what our data is being used for. To apply the criteria proposed in chapter 1 for responsibility to be a response to the other it is essential that ensure that the individuals whose data we are using consent to their data being used in the way that we seek to use it.

As we have seen, privacy is an issue that many of the governmental codes of ethics for Artificial Intelligence do reflect on. The unethical, and potentially at times, unlawful, use of private data is an ongoing topic of concern with few clear or simple answers. This is significantly complicated by the fact that different jurisdictions will often have different laws and regulations with regards to privacy and that digital based companies that are using AI often work internationally (such as within the EU’s GDPR regulations and outside of them). It is tricky to ensure that data is being held securely when said data could easily be being stored halfway across the world in an unrelated jurisdiction.

Bonhoeffer, while rejecting the privatisation of Christian ethics (1949/2009, p 230) does defend the notions of individual autonomy and freedom as being key pillars of a good (in his argument Christian) society. Bonhoeffer also reject any notion that defines the good ‘exclusively as one’s own adherence to principles without any regard for the other person’ (1949/2009, p248). This is the reason that the principle only based ethical systems fail so often. This is a result of a lack of focus on the true subject of ethical consideration - the other person or in other words - our neighbour. Likewise, the need for us to see the protection of privacy as part of the human vocation in the modern world, as suggested in chapter 1, can also assist in our consideration of how to move forward in our dialogue with the legal documents.

At the same time, it is important to recognise that the right to privacy has never been absolute and when we consider the issues of privacy it is not a clear-cut question.

When seeking to apply the questions of responsibility to the issue of privacy we must consider that whatever we do with data must be in the interest of the other, protect privacy, selflessly seek the common good, and be as open and transparent as possible while being prepared to act.

This requires those developing and implementing Artificial Intelligence to ensure that they take vicarious responsible action. It is not sufficient to ensure privacy at the end of a project. Privacy and the considerations regarding how data should be treated, sorted, cleaned, stored etc, must be the first consideration that a developer would need to undertake.

To be responsible requires one to not be reactive or to be undertaking a tick list based on principles. The responsible agent seeks to apply the ethical safeguards as the primary task even before the project is begun (vocation).

Likewise, the responsible agent acts. If there is an issue with privacy it is the duty of the individual, at whatever stage in the process they find themselves to report the issue to supervisors, or if relevant the authorities, or if necessary, even to alert the public such as through news or media. However, in cases of a grave breach of ethics by a company or government the individual must be willing to do whatever is necessary to protect the rights of the other - even to the risk of going public, and facing consequences, when there is the risk of the responsible agent losing their job or worse.

4.4 Bias

Another issue that needs to be faced when considering the ethics of Artificial Intelligence is the risk of bias. This is a theme that we have also seen throughout the various national/ governmental ethics codes that we have considered. One of the most well-known stories relating to AI bias was the case relating to Amazon in which an AI machine did not process job applications in a ‘gender-neutral way’ (Dastin, 2018, NP). As the article by Dastin points out, this took place because of the way in which the model had been trained - using previous resumes and who was hired or not hired as a result. Most of the previous applicants had been men and therefore the machine erroneously learnt that male applicants were therefore preferable to female applicants (Dastin, 2018, NP). As Mo Gawdat said, ‘sadly we are not designing AI to think like a human, we are designing it to think like a man’ (2022, p222).

Bias in Artificial Intelligence is often an inadvertent result of the development process and is a serious risk in many ways. It can be caused by a lack of data, too much data, irresponsibly managed data, and so many other factors involved in the training and development of AI machines.

In terms of ethics and rights such models risk discrimination against people from different ethnic backgrounds, sexualities, genders, gender identities etc. Bias also means that an Artificial Intelligence machine will not be as effective or as objective as it would be otherwise.

A lack of diverse representation within the technological industries is a risk that has, in the development of technology, in the past and in the present resulted in ‘power imbalance in the world and in technology’ (Gebru, 2021, p253). Bias and unconscious bias are found in every person and is a consequence of cultural norms that we inherit by virtue of being a part of a culture. Timnit Gebru argues that ‘AI-Based tools are perpetuating gender’ (2021, p259) and racial stereotypes and that a big part of this is based upon who is developing the relevant machines. She argues that we need to stop viewing science and technology as about ‘finding objective “truths” without taking people’s lived experiences into account’ (Gebru, 2021, p267) and highlights the need for multidisciplinary work to ensure that AI becomes a part of the solution rather than continuing the issues of marginalisation and exclusion (Gebru, 2021, p268). This is a vocational call to the human person because of our obligation to care for the common good of the other.

People need to be able to trust machines if they are going to use them. A lack of trust leads to individuals being unwilling to use the machines and increases the risk of such technology being banned or supressed unnecessarily. Therefore, as the national AI ethics documents recognise, eliminating bias is essential to a responsible development of Artificial Intelligence.

The responsible agent must always consider and act in response to the other, as Bonhoeffer argues, and recognise their own position of privilege and innate bias. Being responsible with the development of Artificial Intelligence recognises that we need to ensure that people from a range of backgrounds and experiences should be involved in the processes of development and implementation. It is no longer sufficient to consider scientific development to always be neutral. Especially given that we have seen through the development of inadvertently bias AI. Our unintentional biases are included and magnified by AI machines. To be responsible therefore means to consider who this technology might impact (the other) and the ways in which this could be harmful to them. Like with privacy - this is not simply something that must be done after the fact but must be asked as a key question at the very start of any project (vocation) and acted upon (vicarious responsible action). Whose experience have I not considered? Are there enough people from minority backgrounds involved in the project? How can I honestly evaluate my biases and work to ensure that they are not included in the development of this project?

The responsible agent must stop and recognise the other - the person who will be impacted by their technology and recognise how they can use this opportunity to selflessly develop the common good.

4.5 Jobs and Employment

One area in which Artificial Intelligence scares individuals the most is the area of jobs and employment. People do fear being replaced by machines and it is something that has already begun to happen in many industries and sectors. In previous revolutions of work, such as the industrial revolution, the developments had required the down skilling of workers. For example, workers went from crafts persons to factory worker. Elliott argues that contrary to past changes to work and employment - the current AI fuelled revolution of work would require the upskilling of the workforce which could potentially lead to vast amounts of unemployment and therefore leaving the smartest and most talented people (or those working directly with AI) with work (Elliott, 2018, NP). This is not limited to those in what are often considered low skilled industries, where we are already seeing technology changing work considerably, for example fast food and customer services, but is something that is far more widespread. Healthcare professionals, teachers, even those working in higher education, can see some of their role replaced, at least in part, by an AI machine. Recently *Chat Open AI* has, in some ways, taken the world by storm; this machine is even capable of writing essays for undergraduates (Hern, 2022, NP) and has been shown to be able to write poetry and even sermons.

Malone, Rus, and Laubacher argue that Artificial Intelligence will continue to dominate our world and that the response that we should give is to 'facilitate the creation of new jobs', that we should be 'matching jobs to job seekers, and providing education, training, and sometimes financial support' (2020, p30) during this transition and indeed this is something that AI can help with. On the other hand, it has been argued that 'AI will be like past technologies, modestly boosting productivity growth and having no effect on the overall number of jobs or unemployment rates' (Atkinson, 2016, p9).

This is a tricky subject and there are no clear cut or easy answers to the many questions surrounding what the future of work will look like.

Work is something that is very important to Bonhoeffer and the entire reformation tradition. Bonhoeffer regarded work as something that is from God and divinely mandated as a part of what it means to be a human person (Green, 2009, p18). The four mandates from God that Bonhoeffer identifies are 'work, marriage, government, and church' (Bonhoeffer, 1949/2009, p68) these can be understood as the realms of vocation. He argues that 'work "in itself" is not divine but work for the sake of Jesus Christ… is divine' (Bonhoeffer, 1949/2009, p69) because work is commanded by God. Bonhoeffer states that, when it comes to work, 'no one can withdraw from this mandate' (Bonhoeffer, 1949/2009, p71) and that through the work of our hands as human persons 'a world should emerge that - knowingly or unknowingly - expresses Christ' (Bonhoeffer, 1949/2009, p71). If the aim of our technological development is to live a life of luxury without any form of work - this is not responsible living because it does not fulfil our call to vocation or to vicarious responsible action.

Responsibility calls us to care for the poor and the other. Especially in this case those, those who through no fault of their own, might lose their jobs because of Artificial Intelligence development and implementation. Work is not an optional extra to the human life and to human persons but is intrinsic to our very dignity. Therefore, the responsible developer of AI will consider the impact of their work on employment and how that impact could be mitigated.

One of the greatest risks in our modern capitalist society is that we seek to maximise profit over and above the common good of the other. However, without giving work dignity in and of itself we lose something of what it means to be human. Technological development is essential, and this too is part of what it means to be human - to be creative and to develop. Yet, a responsible government, or organisation, or individual must pay careful attention to how

Artificial Intelligence is being implemented and consider how to avoid the risk of mass unemployment or the overuse of AI.

It is essential, in responsible Artificial Intelligence development, to consider if the issues that we face, in each instance, even need to be solved by AI in the first place. Does using AI harm someone's dignity (the other)? And as we have said time and time again - what is the purpose of this design? If the design is to benefit humanity and to work for the common good than it is worth considering - if it is simply about profit margins - that is not responsible.

4.6 Climate and Environmental Impact

There are many other ethical issues that we could explore however one of the often-overlooked issues with Artificial Intelligence is the environmental impact of this technology. AI can be a great tool in helping to combat climate change and other ecological issues. Yet, at the same time, studies have shown that 'training a single AI system can emit over 250,000 pounds of carbon dioxide. In fact, the use of AI technology across all sectors produces carbon dioxide emissions at a level comparable to the aviation industry' (Jones and Easterday, 2022, NP). The thing that can be easily forgotten is that when we do something online or via servers the data is always stored somewhere - often in big warehouses that use huge amounts of electricity to cool. Digital appliances can seem like a clean alternative - yet as with many things this is far from straightforward. This is especially true when it comes to big data. As Lucivero points out – big data and the processes used in it 'have a heavy footprint featuring high consumption of non-renewable energy, waste production and CO^2 emissions' (2018, p1010). Big data can be very useful in combatting environmental damage and impacts. However, we are also responsible for the impact of our consumption on the environment – and have an obligation to look for cleaner and more sustainable models of data storage and management for the sake of the other, both present and future.

Responsible Artificial Intelligence must deal with the impact of AI on the environment and take it into consideration in development and implementation. We cannot ignore our responsibility to the environment as responsibility to the environment is a human vocation (Gen 2:15) and includes our responsibility towards the other. An AI machine might be profitable and efficient but if it is going to do harm to the environment than this must be taken into consideration by responsible agents when considering if it is morally right to create a machine. Especially with climate change we have an obligation to work towards the common good. This is an instance in which we are obligated as responsible agents to take up our vicarious responsible action and to act to prevent ecological damage in any way that we can. Sometimes it may be responsible to create an AI machine - AI will hopefully in many ways be a key player in our attempts to reduce the ecological harm that we have caused - but we must recognise the reality of the situation (the real) in which we find ourselves and the reality that creating an AI machine does have an ecological cost and must be thought through fully before such an undertaking is begun.

4.7 Conclusion

There are many ways in which we can apply Bonhoeffer's Ethics of Responsibility to the ethical questions that surround Artificial Intelligence Technology development and implementation. However, the key is that people must recognise the reality of the situation in which they find themselves, take responsibility (vicarious responsible action), even when it is not mandated by the state, to ensure an ethical development of AI technology. To be responsible also involves risk because one must be willing to become guilty (vicarious responsible action) and potentially risk job or reputation to uphold that which is good, right, moral, and ethical. Being responsible means being willing to confront immorality and to fight for what is right - even when no one else is. It is not the responsibility of one company or one individual alone only but rather an ethics of responsibility requires a change of culture and approach where the development of, as Virginia Dignum points out, 'responsible AI requires informed participation of all stakeholders' (2021, p216). Responsible Artificial Intelligence can only be designed selflessly by the participation of all for the common good of all.

Conclusion

This thesis has been seeking to answer if Bonhoeffer's ethics of responsibility can be used as a way forward for Artificial Intelligence ethics. However first we must ask: who is responsible?

It is all too easy to push responsibility onto *the other*. However, as we have seen - Bonhoeffer always points responsibility back toward the person in question. Ultimately one could argue that Bonhoeffer's ethics of responsibility is about each individual person taking hold of responsibility and seeking to act responsibly for the common good (for others). The responsible agent shuns self-serving action and always takes hold of the preferential option for the other. Within such an ethical system "*I"* as an individual take responsibility. The government takes responsibility, the CEO takes responsibility, the implementer takes responsibility, the developer takes responsibility, and ultimately, "*I"* as the consumer take responsibility. In this ethical system no one is excused. Yet how can we apply this practically?

The first thing to note is that an ethics of responsibility is not there to replace the need for professional codes of conduct, for there to be legislation, or for there to be more in-depth considerations on how the ethics of Artificial Intelligence should be viewed or implemented. Instead of seeking to replace professional codes of conduct; our concept of responsibility as expressed in this paper is about the importance of an underlying approach/ culture to ethics that is required in AI. Rather than being about pinning responsibility on one person it is about us collectively taking responsibility for it and changing attitudes to ensure ethical education and involvement at all levels.

As we have seen throughout chapter 2 - Artificial Intelligence technology is nowhere near the point of being like human/ general intelligence. Yet, it is also clear that there are still many ethical questions and concerns that must be responded to and reflected upon. However, what Bonhoeffer's ethics of responsibility does do is point to the necessity of individuals to act and to take up the challenge. If individual responsibility is not at the core of a culture of ethics and legislation, then this is where the risks relating to AI begin to show. If every individual involved takes responsibility it ensures accountability and safety. If people constantly point responsibility elsewhere this is where issues begin.

Ultimately Artificial Intelligence is a tool – it is not an agent that acts on its own behalf. As a result, the question of responsibility lies ultimately in the hands of the user and the developer of said tools. It may be the case that any such tool could be considered too dangerous to use. However, this is a highly unlikely outcome given that it is already in use and does good work in so many different contexts. We should consider AI in a similar way to our consideration of nuclear fusion - it can be incredibly useful and beneficial when used for the creation of power - but when misused can be the cause of much harm.

An Ethic of Responsibility does not alone solve the issues surrounding the difficulties of Artificial Intelligence. However, Bonhoeffer calls us to take an active role in being responsible for the development and implementation of AI, within our own vocation. A responsible ethics of AI must lead to greater transparency and accountability, as many of the documents referenced in chapter 2 point out, but responsibility should compel us to facilitate legislation that encourages responsible and thoughtful practice in the development and implementation of AI machines. In the modern world we cannot see ethics as it was even one hundred years ago. The world has changed in ways that no one could possibly have imagined. However, what Bonhoeffer teaches us is that we must take responsibility for the world *as it is* today and work with the reality of the situation that we face.

'Above all, as Christians… we are called to *'representative action'*… We are to act as if we are responsible for one another and to one another for the sake of Jesus Christ' (Nelson, 2022 B, p55). Bonhoeffer's ethics is a key guiding light to how Christians can respond to the issues presented by Artificial Intelligence technology. However, this principle of responsibility is equally helpful for any person seeking to work towards a safer and more ethical form of Artificial Intelligence. When we all take responsibility for the common good of all humanity, we have the tools to make the world a better place.

This thesis is not the end of the required dialogue. There are many suggestions that need to be considered as the next steps.

Firstly, it is imperative that governments, both nationally and at an international level, develop legislation and oversight of corporations involved in Artificial Intelligence development and implementation. Robust regulation and oversight are prime ways of stopping misuse, bias, and the other issues that we have considered.

Secondly, those involved in Artificial Intelligence development and implementation have a responsibility to humanity. The avoidance of bias and other algorithmic decisions will often lie at the hands of those who develop and implement AI machines. Even with oversight – it is essential that ethics is considered and taken seriously at every level and especially by those involved in hands on application of this technology.

Lastly, all people have a responsibility to be informed and involved. It is not enough for individuals to sit idly by. As we have seen we are all responsible. If we, as ordinary citizens, simply bury our heads in the sand – and fail to be responsible – nothing will change.

Bonhoeffer does not have all the answers to the question of ethics of Artificial Intelligence. However, what his ethics does do is lead us to take responsibility and seek technological development that respects and encourages the dignity of all persons.

Bibliography

Alexander, L, Moore, M. (2020). Deontological Ethics. *Stanford Encyclopedia of Philosophy Website.* Accessed at: plato.stanford.edu/entries/ethics-deontological/ (Accessed 21st of August 2023).

Aoun, J, E. (2017). *Robot Proof: Higher Education in the Age of Artificial Intelligence.* The MIT Press, London, UK.

Aquinas, T. (1485/1991). Summa Theologiae: Question 94. Of the Natural Law. Translated by the Fathers of the English Dominican Province, 1948. *Readings in Moral Theology No7: Natural Law and Theology.* Ed Curran, C, E, and McCormack, R, A. Paulist Press, Mahwah, New Jersey, USA.

Arnett, R, C. (2005). *Dialogic Confession: Bonhoeffer's Rhetoric of Responsibility.* Southern Illinois University Press USA.

Arroyo, C, M. (2002). *The Humanities in the age of technology.* The Catholic University of America Press. Washington D.C., USA.

Ashri, R. (2020). *The AI-Powered Workpace. How Artificial Intelligence, data, and Messaging Platforms are Defining the Future of Work.* Apress, Ragusa, Italy.

Atkinson, R. D. (2016). "It's Going to Kill Us!" and Other Myths about the Future of Artificial Intelligence. *NCSSS Journal*, 21(1), pp. 8–11.

Augsburg Confession. *Book of concord. org.* Accessed at: bookofconcord.org [Accessed 14 Feb 2022]

Bertolini, M, Mezzogori, D, Neroni, M, Zammori, F. (2021). Machine Learning for industrial applications: A comprehensive literature review. *Expert Systems with Applications Volume 175.*

Bethke, A and Schneiderman, K. (2021) AI Ethics Toolkits. *Intel Website.* Accessed at: https://www.intel.com/content/www/us/en/artificial-intelligence/posts/ai-ethics-toolkits.html [Accessed 10 Feb 2021].

Boedy, M. (2017). From Deliberation to Responsibility: Ethics, Invention, and Bonhoeffer in technical communication. *Technical Communication Quarterly, Vol 26, No.2.* Routledge.

Boddington, P (2017). *Towards a code of ethics for Artificial Intelligence*. Springer International Publishing.

Boden, M, A. (2018). *Artificial Intelligence: A Very Short Introduction*. Oxford University Press, New York, USA.

Bonhoeffer, D. (1949/ 2009). *Ethics*. Dietrich Bonhoeffer Works, Volume 6. Translated from the German edition. Edited by: Todt, I, Todt, H, E, Fell, E, and Green, C. English Edition

Edited by Green, C, J. Translated by: Krauss, R, West, C, C, and Scott, D, W. Fortress Press, Minneapolis, USA.

Bostrom, N and Yudkowsky, E. (2014). The ethics of artificial intelligence. *The Cambridge Handbook of Artificial Intelligence*. Edited by Frankish, K and Ramsey, W, M. Cambridge University Press, Cambridge, UK.

Bride, J. (2019). *New Dark Age: Technology and the end of the future.* Verso, London, UK.

Brown, S.G. (2020). Editorial. The Ecumenical Review, 72 (2).

Catholic Church. (2012). *Catechism of the Catholic Church*. Bloomsbury Publishing PLC. London, UK.

Clements. (2022). *Appointments with Bonhoeffer: Personal faith and public responsibility in a fragmented world.* T and T Clark, Dublin, Ireland.

CPCE. (2022). *Christians Speaking of God.* Unpublished Document.

Criddle, C. (2020). Facebook sued over Cambridge Analytica scandal. *BBC News Website.* Accessed at: bbc.co.uk/news/technology-54722362 [Accessed 12 Dec 2022].

Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*. Accessed at: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G [Accessed 28th of Dec 2022].

David Leislie (2019) Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. The Alan Turing Institute. Available at: https://www.turing.ac.uk/sites/default/files/2019-06/understanding_artificial_intelligence_ethics_and_safety.pdf [Accessed 25 Nov 2020].

de Gruchy. (1999). The reception of Bonhoeffer's theology. *The Cambridge Companion to Dietrich Bonhoeffer.* Edited by de Gruchy, J, W. Cambridge University Press, Cambridge, UK (United Kingdom).

Dignum, V. (2021). Responsibility and Artificial Intelligence. *The Oxford Handbook of Ethics of AI.* Edited by: Dubber, M, D, Pasquale, F, and Das, S. Oxford University Press, New York, USA.

Dorries, N. (2022). *Establishing a pro-innovation approach to regulating AI An overview of the UK's emerging approach.* Department for Digital, Culture, Media and Sport. Published by the National AI Strategy, London, UK.

Elliott, S. W. (2018) 'Artificial Intelligence, Robots, and Work: Is This Time Different?', *Issues in Science & Technology*, 35(1). Pp. 40–44.

European Union (EU). (2021). Report/Study. Ethics guidelines for trustworthy AI. *EU Website.* Available at: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai [Accessed 12 May 2021].

Floyd Jr, W, W. (2009). General Editor's Foreword to Dietrich Bonhoeffer Works. *Ethics.* Dietrich Bonhoeffer Works, Volume 6. Edited by: Todt, I, Todt, H, E, Fell, E, and Green, C. English Edition Edited by Green, C, J. Translated by: Krauss, R, West, C, C, and Scott, D, W. Fortress Press, Minneapolis, USA.

Frankish, K and Ramsey W, M. (2014). Introduction. *The Cambridge Handbook of Artificial Intelligence*. Edited by Frankish, K and Ramsey, W, M. Cambridge University Press, Cambridge, UK.

Franklin, S. (2014). History, motivations, and core themes. *The Cambridge Handbook of Artificial Intelligence*. Edited by Frankish, K and Ramsey, W, M. Cambridge University Press, Cambridge, UK.

Gawdat, M. (2022). *Scary Smart: the future of Artificial Intelligence and how you can save our world.* Bluebird Publishers, London, UK.

Gebru, T. (2021). Race and Gender: Data-driven claims about race and gender perpetuate the negative biases of the day. *The Oxford Handbook of Ethics of AI.* Edited by Dubber, M, D, Pasquale, F, Das, S. Oxford University Press, New York, USA.

George, B and Carmichael, C. (2021). *Artificial Intelligence Simplified: Understanding Basic Concepts.* Second Edition. Edited by Mathai, S, S. CSTrends LLP Publishers.

Green, C, J. (2009). Editors Introduction to the English Edition. *Ethics*. Dietrich Bonhoeffer Works, Volume 6. Translated from the German edition. Edited by: Todt, I, Todt, H, E, Fell, E, and Green, C. English Edition Edited by Green, C, J. Translated by: Krauss, R, West, C, C, and Scott, D, W. Fortress Press, Minneapolis, USA.

Green, E. (2020). Sallie McFague and an Ecotheological Response to Artificial Intelligence. *The Ecumenical Review, 72 (2).*

Gregor, B. (2016). The Critique of Religion and Post-Metaphysical Faith: Bonhoeffer's Influence on Ricoeur's Hermeneutics of Religion. *Engaging Bonhoeffer: The Impact and Influence of Bonhoeffer's life and thought.* Edited by Kirkpatrick, M, D. Fortress Press, Minneapolis, USA.

Gunatilleke, J. (2022). *Artificial Intelligence in Healthcare: Unlocking its potential.* N. Janak Gunatilleke, UK.

Gustafson, J, M. (1963). Introduction. *The Responsible Self: An Essay in Christian Moral Philosophy*. Harper and Row Publishers. New York, USA.

Haenlein, M and Kaplan, A. (2019). A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence. *California Management Review, Vol 61 (4).* University of California, USA.

Haidar, Y, A. (2021). Difference between algorithm and model in machine learning. *Linkedin.* Accessed at: https://www.linkedin.com/pulse/difference-between-algorithm-model-machine-learning-yahya-abi-haidar/ [Accessed 21 November 2022].

Hameed, N. (2019). *Multi-class multi-level classification algorithm for skin lesions classification using machine learning techniques.* Faculty of Computing and Engineering, Anglia Ruskin University in partial fulfilment of the requirements of Anglia Ruskin University for the degree of Doctor of Philosophy in Computer Science. Cambridge, UK.

Hao, K. (2020). The USA just released 10 principles that it hopes will make AI safer. *MIT Technology review*. Available at: https://www.technologyreview.com/2020/01/07/130997/ai-regulatory-principles-us-white-house-american-ai-initiatve/ [Accessed 12 May 2021].

Harasta, E. (2014). The Responsibility of Doctrine: Bonhoeffer's Eccleiological Hermeneutics of Dogmatic Theology. *Theology Today Vol. 71 (1)*. Austria.

Harvey, B. (2015). *Taking Hold of the Real: Dietrich Bonhoeffer and the Profound Worldliness of Christianity.* Cascade Books, USA.

Hasselbalch, G. (2021). *Data Ethics of Power: A Human Approach in the Big Data and AI Era.* Edward Elgar Publishing. Cheltenham, UK.

Hern, A. (2022). AI bot ChatGPT stuns academics with essay-writing skills and usability. *The Guardian Online.* Accessed at: https://www.theguardian.com/technology/2022/dec/04/ai-bot-chatgpt-stuns-academics-with-essay-writing-skills-and-usability [Accessed 29 Dec 2022]

Herzfeld, N, L. (2002). *In our Image: Artificial Intelligence and the Human Spirit.* Theology and science. Fortress Press, Minneapolis, USA.

Hughes, G. (1991). The Authority of Christian Tradition and of Natural Law. *Readings in Moral Theology No7: Natural Law and Theology.* Ed Curran, C, E, and McCormack, R, A. Paulist Press, Mahwah, New Jersey, USA.

IBM. (2021). AI Ethics. *IBM Website.* Accessed at: https://www.ibm.com/uk-en/artificial-intelligence/ethics [Accessed 10 Feb 2021].

Jonas, H. (1984). *The Imperative of Responsibility: In search of an Ethics for the Technological Age.* The University of Chicago Press, London, UK.

Jones, E, and Easterday, B. (2022). Artificial Intelligence's Environmental Costs and Promise. *Council on Foreign Relations Website.* Accessed at: https://www.cfr.org/blog/artificial-intelligences-environmental-costs-and-promise#:~:text=Training%20a%20single%20AI%20system,comparable%20to%20the%20aviation%20industry. [Accessed 29 Dec 2022]

Keglar, E, R. (2017). IMAGE ON TITLE. Accessed at: https://www.facebook.com/qchristianorg/photos/we-are-not-to-simply-bandage-the-wounds-of-victims-beneath-the-wheels-of-injusti/10155507019128418/?paipv=0&eav=AfbklGa-tCwqhZeLUrD0lAOBAYlFbbLDOCCuPGt8yqXCtDLqZeaXFHIH2qkd8z_MoMA&_rdr [Accessed 23 Mar 2023]

Kirkpatrick, M, D. (2016) Situations, contexts, and responsibility: Bonhoeffer's Ethics in the thought of Joseph Fletcher, Paul Lehmann, and H. Richard Niebuhr. *Engaging Bonhoeffer:*

The Impact and Influence of Bonhoeffer's life and thought. Edited by Kirkpatrick, M, D. Fortress Press, Minneapolis, USA.

Lennox, J, C. (2020) *2084: Artificial Intelligence and the future of Humanity*. Zondervan Reflective, Chicago, Illinois, USA.

Lewis, D, and Moorkens, J. (2020). A right based approach to trustworthy AI in social media. *Social Media + Society*. Sage.

Loh, W-Y. (2011). Classification and Regression Trees. *Wileys online library*. Joh Wiley and Sons inc. Accessed at: https://wires.onlinelibrary.wiley.com/doi/full/10.1002/widm.8?casa_token=NKIA7GgrbdEAAAAA%3AcS0AINzEh1-_8YVIzPIy-dQrd54-ASf7YNa9YOtiFVlxCG9_xwpk2qiNLqFW0x2zU4shhjBR8_-Uwhc&saml_referrer [Accessed 27 Feb 2023]

Lovin, R, W. (2016). Reinhold Niebuhr and Dietrich Bonhoeffer on Responsibility. *Engaging Bonhoeffer: The Impact and Influence of Bonhoeffer's life and thought.* Edited by Kirkpatrick, M, D. Fortress Press, Minneapolis, USA.

Lutheran/ Roman Catholic Joint Commission. (1980). Ways to Community. *Documentary Supplement*. Accessed at: http://www.christianunity.va/content/unitacristiani/it/dialoghi/sezione-occidentale/luterani/dialogo/documenti-di-dialogo/1980-vie-verso-la-comunione/en.html [Accessed 17 Feb 2021].

Lucivero, F. (2018). Big Data, Big Waste? A Refection on the Environmental Sustainability of Big Data Initiatives. *Sci Eng Ethics* **26**, 1009–1030 (2020). Accessed at: https://doi.org/10.1007/s11948-019-00171-7 [Accessed 6 March 2023]

Malone, T, W, Rus, D, and Laubacher, R. (2020). Research Brief: Artificial Intelligence and the future of work. *MIT Work of the Future*. MIT.

Mealey, A, M. (2009). *The Identity of Christian Morality*. MPG Books Ltd, Cornwall, UK.

Microsoft. (2021). Responsible AI and Microsoft AI principles. *Microsoft Website*. Accessed at: https://www.microsoft.com/en-us/ai/responsible-ai?activetab=pivot1:primaryr6 [Accessed 10 Feb 2021].

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*.

Mueller, J, P, and Massaron, L. (2018). *Artificial Intelligence for dummies*. John Wiley and Sons inc, Hoboken, New Jersey, USA.

Nelson, F, B. (1999). The life of Dietrich Bonhoeffer. *The Cambridge Companion to Dietrich Bonhoeffer*. Edited by de Gruchy, J, W. Cambridge University Press, Cambridge, UK.

Nelson, J. (2022 B). The Rome Call to Artificial Intelligence Ethics Inside the mind of the Machine: How the Church can respond to the ethical challenges presented by AI. *The Trinity*

Conference 2022: A Catholic Response to the Digital Age. Edited by Hayward, H, and Stacey, G. Leeds Trinity University, UK.

Nelson, J. (2022 A). *What can we learn from Dietrich Bonhoeffer's Concept of Responsible Action.* Leeds Trinity PGR Conference 2022. Unpublished.

Nessan, C, L. (2022). *Free in Deed: The heart of Lutheran Ethics.* Fortress Press, Minneapolis, USA.

Niebuhr, R. (1934/ ND). *Moral Man and Immoral Society*. Kessinger's Legacy Reprints. Charles Scribner's Sons. New York, USA.

Niebuhr, R. (1963). *The Responsible Self: An Essay in Christian Moral Philosophy*. Harper and Row Publishers. New York, USA.

Nissen, U, B. (2011 A). Letting Reality Become Real: On mystery and reality in Dietrich Bonhoeffer's Ethics. *The Journal of Religious Ethics Vol 39 (2).*

Nissen, U, B. (2011 B). 'Responsibility and Responsiveness: reflections on the communicative dimension of responsibility. *Neue Zeitschrift fur systematische Theologie und Religionshphilosophie*, 53 (1).

OpenAI. (2023) ChatGPT. *OpenAI websire.* Accessed at: openai.com/blog/chatgpt/ [Accessed 27 Feb 2023].

Paulson, S, D. (2011). *Doing Theology: Lutheran Theology*. T & T Clark International. London, UK.

Rai, A. (2019). Explainable AI: from black box to glass box. *Journal of the Academy of Marketing Science 2020: 48:137- 141.*

Rasmussen, L. (1999). *The Cambridge Companion to Dietrich Bonhoeffer.* Edited by de Gruchy, J, W. Cambridge University Press, Cambridge, UK.

Rassmussen, L. (2009). Dietrich and Karl-Friedrich Bonhoeffer. The Brothers Bonhoeffer on science, morality, and technology. *Zygon Vol 44, no 1.*

Reed, D, E. (2020). *The Limit of Responsibility: Dietrich Bonhoeffer's Ethics for a Globalizing Era.* Series Editors Brock, B and Parsons, S, F. T & T Clark Enquiries in Theological Ethics. Bloomsbury Publishing PLC, London, UK.

Reuters (2018). Amazon ditched recruiting tool that favoured men for technical jobs. *The Guardian (online).* Available at: https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine [Accessed 20 Nov 2019].

Ricoeur, P. (2007). *Reflections on The Just*. Translated by Pellauer, D. The University of Chicago Press, Chicago, USA.

Rimell, W. (2022). Autonomous cars to become legal in UK "within the next year": New legislation that will allow autonomous driving on motorways to be debated by government. *Autocar Website.* Accessed at: https://www.autocar.co.uk/car-news/consumer/autonomous-cars-become-legal-uk-within-next-year#:~:text=The%20first%20cars%20that%20can,autonomous%20vehicles%20planned%20from%202025 [Accessed 12 Dec 2022].

Roszak, T. (1986). *The Cult of Information: The Folklore of Computers and the True Art of Thinking*. Paladin Grafton Books, London, UK.

Rudin, C. (2019). *Stop Explaining Black Box Machine Learning models for High stakes decisions and use interperatble models instead.* Duke University.

Sankaran, V. (2023). Microsoft lays off team responsible for ethical AI development. *The Independent.* Accessed at: www.independent.co.uk/tech/microsoft-layoff-ai-ethics-team-b2300237.html [Accessed 22 Mar 2023].

Schneier, B. (2018). *Click here to kill everybody: Security and survival in a hyper-connected world.* W.W. Norton Company, New York, USA.

Schuurman, D.C (2019). Artificial Intelligence: Discerning a Christian Response. *Perspectives on Science and Christian Faith, 71(2).* Available at: https://link.gale.com/apps/doc/A595353613/AONE?u=tasc&sid=AONE&xid=bbf85a16 [Accessed 02 Dec 2020].

Shukla, L. (2019). Designing Your Neural Networks. *KD Nuggets*. Accessed at: kdnuggets.com/2019/11/designing-neural-networks.html [Accessed 13 Jan 2021].

Smith, B. W. (2020). Ethics of Artificial Intelligence in transport. *The Oxford Handbook of Ethics of AI.* Edited by: Dubber, M, D, Pasquale, F, and Das, S. Oxford University Press, New York, USA.

Sun, R. (2014). Connectionism and Neural Networks. *The Cambridge Handbook of Artificial Intelligence*. Edited by Frankish, K and Ramsey, W, M. Cambridge University Press, Cambridge, UK.

Taulli, T. (2019). *Artificial Intelligence Basics: A Non-Technical introduction.* Apress, New York, USA.

Taulli, T. (2020). Deep Learning: What You Need To Know. *Forbes*. Available at: https://www.forbes.com/sites/tomtaulli/2020/03/27/deep-learning-what-you-need-to-know/?sh=fd052176f30b [Accessed 8 Jan 2021].

Turing, A. M. (1950). Computing, Machinery, and Intelligence. Mind, Volume 49.

UK Government. (2019). Guidance: Understanding artificial intelligence ethics and safety. *Gov.uk website.* Accessed at: https://www.gov.uk/guidance/understanding-artificial-intelligence-ethics-and-safety [Accessed 17 Feb 2021].

Vatican II. (1965). Gaudium et Spes. *Vatican Council II. 1988 Revised Edition*. Ed. Flannery, A. Costello Publishing Company, New York USA.

Veith, G, E Jr. (2021). *The Spirituality of the Cross: The way of the First Evangelicals*. Third Edition. Concordia Publishing House, St Louis, USA.

Weber, M. (1917/2004). *The Vocation Lectures: 'Science as a Vocation' 'Politics as a Vocation'*. Edited by Owen, D. Strong, T,B. Translated by Livingstone, R. Hackett Publishing Company, Indianapolis, USA.

Wingren, G. (2004). *Luther on Vocation.* Translated by Rasmussen, C, C. Wipf and Stock Publishers, Oregon, USA.

Wyss Institute. (2022). RoboBees: Autonomous Flying Microrobots. *Wyss Institute*. Accessed at: wyss.harvard.edu/technology/robobees-autonomous-flying-microrobots/ [Accessed 30 Nov 2022].

Young, S. (2020). The Future of Farming: Artificial Intelligence and Agriculture. *Harvard International Review*. Accessed at: hir.harvard.edu/the-future-of-farming-artificial-intelligence-and-agriculture/ [Accessed 30 Nov 2022].