Human-Artificial Intelligence Symbiosis: the Possibility of Moral Augmentation

Miletić, Tomislav

UNIVERSITY OF RIJEKA

FACULTY OF HUMANITIES AND SOCIAL SCIENCES

DEPARTMENT OF PHILOSOPHY

Tomislav Miletić

Human-Artificial Intelligence Symbiosis: the Possibility of Moral Augmentation

DOCTORAL THESIS

Rijeka, 2021

UNIVERSITY OF RIJEKA

FACULTY OF HUMANITIES AND SOCIAL SCIENCES

DEPARTMENT OF PHILOSOPHY

Tomislav Miletić

Human-Artificial Intelligence Symbiosis: the Possibility of Moral Augmentation

DOCTORAL THESIS

Mentor: prof. dr. sc. Elvio Baccarini

Komentor: dr.sc. Dražen Brščić

Rijeka, 2021

Mentor rada: prof. dr. sc. Elvio Baccarini

Doktorski rad obranjen je dana 21. siječnja 2021. u Rijeci,

pred povjerenstvom u sastavu:

1. doc. dr. sc. Nebojša Zelič, predsjednik, Filozofski fakultet u Rijeci
2. izv. prof. dr. sc. Luca Malatesti, član, Filozofski fakultet u Rijeci
3. prof. dr. sc. Darko Polšek, član, Filozofski fakultet u Zagrebu

Acknowledgment

I acknowledge that this doctoral thesis has been concluded during my stay in HRI Laboratory, Department of Social Informatics, Graduate School of Informatics, Kyoto University, Kyoto, Japan as the International Research Fellow (PE19056) of Japan Society for the Promotion of Science, supported by the "FY2019 JSPS Postdoctoral Fellowship for Research in Japan (Short-term)".

Zahvala

Ova je doktorska radnja rezultat mog osobnog i istraživačkog puta unazad proteklog desetljeća. Iako je, kroz godine, doprinos mnogih utjecao na njezino stvaranje, njezino se konačno oblikovanje bez potpore nekolicine osoba ne bi ostvarilo. U tom vidu se, želim zahvaliti svom mentoru, Elviju, na njegovom akademskom vodstvu, usrdnoj pomoći i pravom prijateljstvu. S velikim zadovoljstvom mogu reći da sam bio podučen i inspiriran od učitelja i ljubitelja mudrosti – iskustvo i primjer za nasljedovati! Jednako tako, ostajem zahvalan i svom komentoru, Draženu, bez kojega nikada ne bih došao u Kyoto, Japan (svoj dugogodišnji san!) i u miru, uspješno, priveo pisanje ovog rada svome kraju te ostvario uspješnu međunarodnu suradnju u polju ljudsko-robotske interakcije. Stojim zahvalan i svojoj obitelji, ponajviše roditeljima, bratu i sestri s njihovim obiteljima, koji su mi pružali moralnu motivaciju i materijalno dobro. Jednako tako, s velikom se radošću zahvaljujem mojoj braći u vjeri, istinskim prijateljima. Prvenstveno se velikim slovima zahvaljujem NIKICI JURIĆU koji mi je pružio najizdašniju, iznimnu, pomoć u dva ključna časa moje priprave za Japan, te mi tako ne samo iskazao svjedočanstvo prave Kršćanske darežljivosti već i istinskog prijateljstva! Zahvaljujem se i Josipu Petehu, vjernom kumpanjonu čije je duhovno i materijalno dobročinstvo na meni ostvarilo ono što mu ime obznanjuje - "Gospodin nadopunja"! Hvala i mojim dragim prijateljima Mariu i Petru koji me nisu zaboravili kako s potporom tako i s darežljivom rukom! Hvala i tebi Kristijane, istinski čovječe, prokušani ratniče! Hvala ti od svega srca na prijateljstvu i radosti koju si uvijek sa mnom dijelio, na „hidden power" na koju si me uvijek upućivao, snagu s kojom sve u životu postaje „lightweights baby!" Konačno, hvala tebi Martina, moj ostvareni snu! I na samome kraju, hvala tebi Vječni Duše. Nitko od nas ne može znati ima li tebe uopće. Ali ja stojim zahvalan, što me odgovor na to pitanje više ne zanima. Radije, „all that matters is that today, two stood against many" and that „in dawn the beauty reigns and the way is clearer".

Abstract

This thesis posits the hypothesis that the formation of a symbiotic relationship between a child and its AI companion can create a realistically applicable, socially acceptable, fine-tuned, moral augmentation aligned with the child's moral development. The first chapter presents the concept of moral augmentation, introduces the idea of symbiosis, and delineates the proposal's philosophical and technical differences from existing moral enhancement approaches. The second chapter grounds the discussion of AI capacities on realistic technical possibilities and shows why these machine limitations lead to ethical symbiotic partnerships. The third chapter explicates the necessary ethical requirements for the design of artificial ethical agents and shows how these design requirements constitute ethical means by which the AI companion accomplishes moral augmentation. The fourth chapter expounds on the concept of Human-Artificial Intelligence Symbiosis. The fifth chapter proposes the symbiotic AI companion as the ethical artificial intelligence means towards the accomplishment of moral augmentation. First, the companion paradigm is provided, second, the ethical capacities by which the AI companion aims to achieve moral augmentation with the human child are elucidated, third the question of companion distribution in a democratic society is explored.

Keywords

AI ethics, AI companions, Human-AI symbiosis, Moral augmentation,

Prošireni sažetak

Ova teza postavlja hipotezu kako stvaranje simbiotskog odnosa između djeteta i njegovog AI suputnika može stvoriti realno primjenjivu, lako prihvatljivu, fino podešenu, moralnu nadogradnju usklađenu s djetetovim moralnim razvojem.

Prvo poglavlje predstavlja koncept moralne nadogradnje, uvodi ideju simbioze i ocrtava filozofske i tehničke razlike prijedloga od postojećih pristupa moralnom poboljšanju: temeljne motivacije, konačne ciljeve moralnog poboljšanja i tehnološka sredstva za postizanje tih ciljeva.

Drugo poglavlje utemeljuje raspravu o sposobnostima umjetne inteligencije na realističnim tehničkim mogućnostima i razlaže zašto strojna ograničenja dovode do simbiotskog partnerstva. Rasprava se usmjerava na metode strojnog učenja: učenje pod nadzorom, učenje bez nadzora, učenje podrškom.

Treće poglavlje objašnjava potrebne etičke zahtjeve u dizajnu umjetnog etičkog agenta i pokazuje kako etički principi konstituiraju etička sredstva pomoću kojih AI suputnik postiže moralnu nadogradnju. Razlaže se i polje etike umjetne inteligencije, te strojna etika i robotska etika. Analizira se osnovne etičke principe umjetne inteligencije: robustnost, poštivanje ljudske autonomije, objašnjivost i strojna pravednost. Ovdje se također pojašnjava struktura etičkih modela koji upravljaju etičkim odlukama umjetnog agenta te se razmatra problem pouzdanosti umjetnih inteligencija i ljudske odgovornosti u etičkim odlukama.

Četvrto poglavlje objašnjava koncept simbioze čovjeka i umjetne inteligencije. Prvo, izlažu se razlozi potrebe simbiotskog odnosa. Drugo, definira se simbiotski odnos i predlažu dvije verzije: jaka i slaba. Ilustriraju se simbiotska parterstva i analiziraju vitalne točke koncepta: autonomija, utjelovljenje, strojna ekspertiza, motivacijska potpora, pouzdanost.

Peto poglavlje predlaže simbiotskog pratioca umjetne inteligencije kao društveno prihvatljivu etičku umjetnu inteligenciju za postizanje moralne nadogradnje. Prvo, izlaže se paradigma suputnika, pružaju se primjeri i analiziraju osnovne karakteristike – vjernost, transparentnost, fina podešenost, poštivanje moralne autonomije. Drugo, razjašnjavaju se etički kapaciteti kojima AI pratilac želi postići moralnu nadogradnju kod ljudskog djeteta. Oni uključuju etičke modele kojima suputnik donosi valjane etičke savjete, te afektivno računarstvo, metode umjetne inteligencije kojima se prepoznaje, intepretira i simulira ljudske emocije. Ovim metodama suputnik cilja ostvariti moralnu motivaciju i emotivnu stabilnost. Konačno, istražuje se problematika distribucije suputnika u demokratskom društvu. Ovdje pokazujem na koji način distribucija suputnika umjetne inteligencije može pogoršati postojeće asimetrije moći u društvu i ugroziti demokratske procese. Kao moguće rješenje predlažem usvajanje regulatornog pristupa.

Ključne riječi

etika UI suputnici UI moralna nadogradnja simbioza ljudi i UI

Contents

“All we have to decide is what to do with the time that is given us.”

J.R.R. Tolkien, The Fellowship of the Ring

INTRODUCTION

The prospect of moral enhancement aims to achieve a positive-sum net increase in human morality beyond what is usually considered normal for human beings. Since the idea's original proposal (Douglas, 2008; Persson & Savulescu, 2008) several fundamental points have crystalized. These include the issue of unwillingness or the rejection of moral enhancement due to ethically unacceptable or ineffective means, the question of children enhancement, and the very capacity of technical means to achieve a person-specific, fine-tuned, moral enhancement. For this work, I summarize these as moral autonomy, unwillingness, children enhancement, and fine-tuning.

However, the proposal of moral enhancement has recently been cast into doubt with strong rebuttals engaging both its philosophical tenets and technical means (Buchanan & Powell, 2018; Crockett, 2014; Dubljević & Racine, 2017; Paulo & Bublitz, 2019; Wiseman, 2016). In lieu of these discussions the most recent alternative has been formed - that of artificial moral enhancement. This approach aims to be more efficient and less controversial than previous proposals of moral bioenhancements as it aims to utilize the means of artificial intelligence advisors to enhance human morality (Giubilini & Savulescu 2018; Savulescu & Maslen, 2015).

Unfortunately, in its current state, the artificial moral enhancement proposal does not explicate on the ethical nature of the artificial advisors, does not engage the question of children enhancement, does not propose attunement with early developmental phases, and does not provide means by which to improve acceptance or tackle the problem of unwillingness.

The aim of this thesis is then to 1) ground the concept of moral augmentation as a philosophically prudent, technically achievable, and socially acceptable proposal of technologically empowered means for moral improvement in the 21st century, 2) define and analyze the concept of human-AI

symbiosis and 3) propose the symbiotic relationship between a child and an AI companion as the acceptable and finely-tuned moral augmentation means aligned with the child's moral development.

To establish these goals, this work utilizes the methods of philosophical and conceptual analysis, deductive and inductive argumentation, and reflective equilibrium. Relying on empirical evidence as argumentative support will be provided. Vital intuitions will be clarified and appealed to. Thought experiments will be used to exemplify the conclusions of ethical analysis and to narratively describe symbiotic models. Phenomenological introspection will be utilized to represent important experiential dimensions of the symbiotic relationship. Lastly, ethical narration will be used to initiate important ethical reflections.

To properly introduce the reader to the main body of the argument, in this introduction I first present the plan of the thesis. After it, I explicate the current state of the debate and provide reasons why its philosophical development resulted in the most recent proposal of artificial moral enhancement. I will then analyze the vital points of this latest approach and propose a differential improvement with the concept of symbiotic moral augmentation.

The Plan of the Thesis

In the first chapter, I explicate the concept of moral augmentation. Initially, in the first two subchapters, it will be shown how the proposal of moral augmentation differs from the concept of moral enhancement. Here I present three important differences: underlying motivations, the final goals of the enhancement procedures, and the technological means to achieve these goals. From these, I seamlessly arrive at the concept of human-AI symbiosis, which I introduce in the "Symbiosis

and Augmentation" subchapter. Here I first introduce the history of the concept and show how it closely relates to contemporary AI technologies, which makes it a prime candidate of fine-tuned, ethical, artificial intelligence means of moral augmentation. Second, I provide the initial analysis of symbiotic systems which I compare to systems of distributed and systems of extended agency. Here I argue that symbiotic systems share characteristics of both - they are collaborative, distributed systems with a finely-tuned, close coupling between its constitutive members. Finally, I introduce why, in principle, this kind of fine-tuned coupling can build familiarity and trustworthiness which can tackle the issue of unwillingness (i.e. augmentation acceptance) in moral augmentation.

In the second chapter, I exemplify what kind of fine-tuned operations can be, in general, achieved with the technical limitations and opportunities of the current and close-future level of AI technologies. The purpose of this chapter is to ground all the subsequent discussion on AI capacities within these realistically achievable possibilities. Thus, in the first three subchapters, "Supervised Learning", "Unsupervised Learning" and Reinforcement Learning" I point to the strengths and weaknesses of each method and show how these allow machines to achieve super-human performance in computational tasks. I also clarify what are the intrinsic machine limitations of these methods, and why these necessarily make AI ethics a design requirement for AI systems, and how these ethical demands naturally lead to the development of symbiotic systems.

In the third chapter, I expound on the reasons why ethical AI means by which we aim to achieve moral augmentation ethically matter. As such, this part serves to introduce how exactly machines bring about ethical decisions, and what necessary ethical requirements have to be considered when designing ethical machines. To do so I have to first, introduce the field of AI ethics as a sub-field of applied ethics, subdivided into two distinctive but often overlapping fields of research:

Machine ethics and Roboethics. Second, I exemplify what basic ethical principles, what ethical building blocks, all ethical models guiding the behavior of artificial agents have to consider or implement within their operations. Inspired by the European Commissions' Ethical guidelines for Trustworthy AI I analyze the principles of technical robustness, human autonomy, explicability, and fairness. To each, I dedicate a distinct section.

In the next subchapter, I discuss the critical question of AI bias as I expose the concepts of AI reliability and human responsibility. Here I make an argument, that due to the machine's inherent limitations machines can only be evaluated as epistemically reliable, not epistemically responsible and I show why this conclusion speaks in favor of the symbiotic relationship.

In the fifth, final subchapter I exemplify how these principles are utilized to build some of the ethical models with which the machine evaluates ethical scenarios and produces ethical decisions. These include the top-down models utilizing deontological and consequentialist theories, the bottom-up virtue ethics models, and hybrid models utilizing some combination of both.

In the fourth chapter, I set forth to elaborate the concept of human-AI symbiosis. First, I summarize the reasons why the substantial technical limitations and ethical requirements and demands naturally lead to the need for symbiotic partnership in real-life, human-machine collaborations.

Next, I define the human-AI symbiotic relationship and present two symbiotic versions, a strong and weak version and show what kind of fine-tuning is necessary to achieve a symbiotic relationship that respects and augments human autonomy. In the third subchapter „Symbiotic partnership“ I exemplify some of the constitutive characteristics of a symbiotic partnership by utilizing real-life examples of humans and police/military dog partnerships. Derived from these

examples, I analyze vital symbiotic points: autonomy, embodiment, machine expertise, motivational support, trustworthiness.

In the fifth chapter, I explicate the concept of the symbiotic AI companion - a symbiotic artificial agent that achieves moral augmentation with a human child. Here I show how the formation of a symbiotic system with a human child is capable of effectively tackling the issue of unwillingness, fine-tuning, and vital developmental phases to accomplish practical moral augmentation.

I first define and exemplify the companion paradigm. I do this in the first two subchapters where I, first, provide narrative examples of role-play companions and reflect on their moral characters. From there I proceed to analyze the traits of a companion, which improve acceptance: fidelity, transparency, fine-tuning, respect for moral pluralism.

In the third subchapter, „Companion's capacities“ I show what sort of minimal capacities should the AI companion need to have to establish moral augmentation. These are then the, operationally understood, means of moral augmentation. First, I focus on affective computing. Here I show how the regulation of emotional states by the AI companion bears a direct beneficial impact on moral decision-making and moral development. I also exemplify the limits of the current state-of-the-art affective technologies and show why the introduction of moral augmentation in early childhood is an optimal point, both from the viewpoint of acceptance and fine-tuning.

Second, I propose a tri-partite ethical model that is capable of generating valid ethical advice as it takes into consideration both the child's inputs, the parent/caretaker inputs, and basic moral constraints. I expound on the ethical operations of all the three models and explore how the companion utilizes the method of reflective equilibrium to produce valid and fine-tuned moral advice aligned with the child’s developmental phases. Here I also warn of the danger of moral de-

skilling. I analyze to what kinds of generated advice is this concern related to and propose to mitigate moral skill diminishment through open-ended advice which fosters the practice of moral autonomy on the part of the human agent.

In the last chapter, that of "Companion distribution" I show how a wide-spread distribution of AI companions might exacerbate power asymmetries and fracture democratic processes. I analyze three expectable policy approaches, the affirmative approach, the prohibitory approach, and the regulatory approach, and reflect on companion distribution within each. To engage the issue of power asymmetry, and warrant the companion's distribution beneficial outcome I propose the adoption of the regulatory approach.

I conclude my work within the style of narrative ethics.

Moral enhancement: an overview

> "Humane technology starts with an honest appraisal of human nature. We need to do the uncomfortable thing of looking more closely at ourselves." —Tristan Harris

At the writing of this text, the concept of moral enhancement (Douglas, 2008; Perrson & Savulescu, 2008), is over a decade old. Originally stemming from the transhumanist, enhancement, circles, and the initial idea was rather straightforward. If we want to improve humanity's overall well-being and the chances of survival on our planet, then cognitively and physically enhancing humanity will not be enough. Instead, we require radical moral improvement as the significant, existential risks of our times require moral and not just cognitively improved humanity. In essence, moral enhancement is about the free choice of a rational individual to obtain moral capacities higher than the ones provided by biology and to utilize those capacities to guide humanity's future towards better ends.

The goal is, then, to achieve a sum net increase „in the way in which we act or reflect morally" (Raus et al., 2014). Similar to other types of enhancements, the levels of improvement in moral enhancement are to go beyond what is usually considered normal for human beings. Various technological solutions were proposed as a means to this goal, all sharing a universal requirement. Negative outcomes or adverse side effects for the rational individual (human adult) who undertakes the procedure in her free will are to be avoided (Douglas, 2008; Douglas, 2013).

As can be expected, the moral enhancement proposal has, throughout the years, fostered many ethical debates on the desirability, validity, and value of such a project. Since the topic's initial conception, several of these critical issues have come to the forefront. These include, first, the issue of unwillingness or the rejection of moral enhancement due to ethically unacceptable or ineffective means. Second, the question of children's enhancement and third, the very capacity to achieve a person-specific, fine-tuned moral enhancement.[1] Summarized, they are moral autonomy, unwillingness, children enhancement, and fine-tuning. In debating these critical issues, distinguished scholars have, with time, taken residue in two distinctive camps, the traditional and the biomedical camp. These have remained to be the stable viewpoints for the debate over different moral enhancement proposals and as such will be here taken into consideration.

[1] Here, I put aside the notion of enhancement distribution. If the procedure proves valid and effective, the next socio-political question is then one of a wide-spread, fair distribution of enhancement. This question is important to have in mind for a realistic proposal prospect of improvement. I just do not focus on it here, but rather in the last part of the discussion.

The traditional approach

The first camp aims to achieve the goal of moral enhancement through more traditional, deliberative means such as education or moral training (Harris, 2010) rather than biomedical methods. Traditional means to moral enhancement primarily include moral education, personal moral training, and upbringing. Secondary, they may include different social and political means which aim to amplify moral deliberation, emotional control, and moral motivation.

The benefits of the traditional approach mainly lie in its safe and proven validity. Especially in education and upbringing, the traditional means to accomplish moral improvement have been refined and tested throughout many decades. They are capable of utilizing the full advantage of scientific discoveries in developmental and cognitive psychology, education sciences, and medicine. They do not intrinsically jeopardize one's autonomy and individual liberty or withhold the capacity to instigate radical changes in personal identity or biology. Traditional means are also recognized as methods that can preserve and enhance personal autonomy, rational free choice, and the development of one's moral and social identity. Their use bears no strong incentive to lock-in a specific set of moral standards - which might be the case when utilizing biomedical means. Moreover, traditional means pose no grave risk for one's identity or biology. As humans generally eschew radical changes in fundamental aspects of their (moral) character, traditional means can be readily accepted. This point is especially vital concerning children enhancement as traditional means, exemplified through education, withhold no eugenic-like danger. Instead, they support the development of moral character and the convergence of different moral influences – educational institutions, parents/caretakers, society as a whole.

Crucially, traditional means are also intrinsically respectful to the importance of moral development, especially with regards to its critical periods. These sensitive threshold periods are essential for the development of neurobiological systems that facilitate the growth of emotional and social intelligence, empathy, and higher cognitive functioning. As these directly impact self-regulation and autonomous agency, it is vital to safeguard them. Also, since the status of neurobiological development formed in this crucial developmental phase is, once established, hard if not outright impossible to change later on in life (despite the brain's plasticity) early childhood points to be the optimal starting position for moral enhancement (Christen & Narvaez, 2012).

Still, the major disadvantage of the traditional approach and the reason why the opposite, biomedical approach harbored such excellent traction, ever since the field's initial conception, lies in its purported ineffectiveness. The (in)effectiveness includes the range of speed, depth, and scope of enhancement – crucial elements that the biomedical approach aims to improve (Persson & Savulescu, 2008). The biomedical approach aims to solve this ineffectiveness by directly improving the capacity of human biology to think, act, and experience morality. Means to accomplish such a goal can include genetic and cellular interventions, and mostly, as the past decade of the debate has exemplified - neuropharmacology.

The biomedical approach

Although initially highly promising, the biomedical approach has generated many objections, both philosophical and empirical. These include, first, the production of unintended and unwanted side effects, which can result in long-lasting changes in one's identity. Such changes are also capable of creating obstacles to moral enhancement acceptance, even if proven beneficial and useful. Second, even if effective, the biomedical means might inadvertently hinder the development of

moral character and moral capacities by providing a quick-fix solution to moral problems (i.e., a moral pill). Here the adjustment of complicated moral matters would not be achieved by moral deliberation, trained moral responses, or even public policy but through direct biological change. Moreover, such direct adjustments might easily create a one size fits all solution which can diminish the plurality of moral views and the practice of moral autonomy. Finally, creating fundamental changes in human character might prove to be a hindrance rather than the motivation for moral enhancement acceptance (Douglas, 2013). Such obstacles to enhancement can be further exacerbated with regards to the utilization of biomedical means on children. The biggest ethical concerns, here, connect with eugenic fears – the question of who gets to choose what biological changes we opt for, and what those changes will be. In the case of children, the question is, will parents decide or the state's scientists?

Also, the connection between moral development and moral enhancement has been mainly neglected due to the debate's focus on the voluntary choice, the free choice of the individual rational agent (i.e., an adult citizen) to pursue moral enhancement procedures (Douglas, 2008).[2] And this simple fact, of human biological development, holds possibly the most significant impact on the practical possibility of moral enhancement. The reason behind this conclusion lies in the "critical periods," particular types of sensitive threshold periods crucially essential for the development of critical neurobiological systems. These facilitate the growth of emotional and social intelligence, empathy, and higher cognitive functioning, the possibility for self-regulation,

[2] This decision aims to evade many ethical dangers which can otherwise be produced, especially with regard to one's autonomy and identity. These can include, the danger of moral compulsion, diminishment of individual rights and personal autonomy, the forced change of identity, threats to social order and democracy and the danger of eugenics. For a summarized list of arguments see (Specket et al., 2014).

and autonomous agency.[3] As Narvaez has remarkably pointed out, there are some developmental phases in human neural biology that can never, if gone wrong, later on in life, be ameliorated nor enhanced. Conclusively, this makes early childhood an optimal starting point for moral enhancement:

> „How often do you go into a panic? A rage? Feel anxious? If something like this happens to you regularly or frequently, your 'present' may be governed by things that happened in your past. When you go into such a brain state, you usually cannot perceive accurately what is happening before your eyes. Instead, old memories take over and affect what you 'see.'…When brains are under cared for, they become more stress-reactive and subject to dominance by our survival systems—fear, panic, and rage. One's free will can be undermined by how the brain-body complex was shaped in sensitive periods, particularly in early life. " (Narvaez, 2014)

Rebuttals

However, after propelling the vision of moral enhancement for almost a decade, the biomedical approach, especially in its staple and dominant neuropharmacological iteration, has recently been cast into doubt with strong rebuttals (Buchanan & Powell, 2018; Crockett, 2014; Wiseman, 2016). These put into question the very feasibility of the biomedical option for moral enhancement

[3] "Much of early experience involves the co-construction of brain circuitries through interactions with caregivers that have a direct bearing on moral functioning" (Christen & Narvaez, 2012, p. 2). "Resiliency research in developmental psychology indicates that the brain is plastic beyond early childhood, though some thresholds and systems are established in early life and may be hard if not impossible to change. In any case, early childhood is probably the optimal starting point for moral enhancement" (Christen & Narvaez, 2012, p. 5).

(Dubljević & Racine, 2017). The most important include the blunt ineffectiveness of the current state of biomedical technologies to obtain the goal of moral enhancement in a person-specific fine-tuned way (Douglas, 2013). Also, they risk causing adverse side-effects, especially in neuropharmacological manipulations (Therbeck & Chesterman, 2014). For instance, they can inadvertently create undesired and long-lasting structural changes within the neurotransmitter system's vastly complex and interconnected system. To name a few, Oxytocin was found to promote trust, but only in the in-group. At the same time, with the out-group members of society, it can decrease cooperation and selectively promote ethnocentrism, favoritism, and parochialism. As such, Oxytocin has a „nudge" potential but only with those who are already disposed towards prosocial behavior or empathy (Wiseman, 2016). Additionally, Beta-blockers were found to decrease racism but also blunt all emotional responses, which put their practical usefulness in general doubt. SSRIs (Selective Serotonin Reactive Inhibitors) reduce (reactive) aggression, but have serious side-effects, including an increased risk of suicide (Wiseman, 2016) and increase pre-mediated aggression. Also, purposely manipulating bodily Serotonin levels can lead to harmful consequences – a vast and validated problem of neuropharmacology in general (Terbeck & Chesterman, 2014). Molly Crocket warns,

> "Most neurotransmitters serve multiple functions and are found in many different brain regions…serotonin plays a role in a variety of other processes including (but not limited to) learning, emotion, vision, sexual behavior, appetite, sleep, pain and memory, and there are at least 17 different types of serotonin receptors that produce distinct effects on neurotransmission Thus, interventions that affect moral behavior by globally altering neurotransmitter function may have undesirable side effects, and these should be considered when weighing the costs and benefits of the intervention". (Crockett, 2014).

The harm of adverse side-effects is directly imposing high-safety concerns upon such technologies. Still, it is also showcasing the inability to obtain the person-specific fine-tuned (Douglas, 2013) goal of moral enhancement.[4] Additionally, the issue of fine-tuning is also reflected in the inadequacy of scientific models[5] upon which moral enhancement researchers tend to rely on their ethical assessments. Namely, these laboratory models focus mostly on investigating moral behavior inside constrained experimental frameworks. These are neither contextualized nor experienced from a first-person perspective[6] and often do investigate moral decision making and action within real-life scenarios.[7]

> "Moral psychology research today focuses mostly on moral judgment in constrained experimental tasks and does little to investigate the moral decision-making and action that occurs on a moment-to-moment basis, where perceptions interact with situations to promote shifting moral mindsets." (Narvaez, 2014)

[4] „That what is necessary for moral enhancement is the fine tuning of certain emotions in a person-specific way that is sensitive to prevailing circumstances, not the wholesale elimination of emotions at a population level." (Douglas, 2013).

[5] Needless to say the problem might be doomed to fail from the start. "If that turns out to be the case, then no singular model of moral judgment can possibly specify a particular brain process in every human brain to target for some enhancement, and the issue is not related to technology at all. Further empirical work might provide answers to these vexing possibilities." (Dubljević & Racine, 2017, note 48)

[6] "...most of the science upon which moral enhancement enthusiasts draw is conducted either using Ivy League students or mice" (Wiseman, 2016, p. 117).

[7] Neurological underpinning of other-oriented perspective differs from that of self-oriented perspective. In self-oriented perspective merely seeing other's facial expression and decoding the facial signals, easily fosters direct motivational tendencies, for instance the flight or fight responses (Kim & Son, 2015). This does not happen with other-oriented perspectives. (Coplan, 2011)

Finally, the vexing issue of unwillingness remains unsolved. This is the reality of how there are humans who do not wish to be morally enhanced. As shown in empirical findings (Douglas, 2013)[8] , morally significant traits such as empathy and kindness are often least willed to be enhanced. This would put into doubt the goal of a globally distributed moral enhancement even if the technologies were applicable which, again, puts a significant focus on the value of development and education. As Wiseman concludes:

> „Narcissism, moral cynicism, unwillingness to reflect upon the plight of those who are suffering, and the like are highly complicated phenomena; none of these concerns are monocausal or even predominantly biologically based issues, to begin with. What sort of technology might be advanced that would help persons be more inclined to think about issues of poverty and inequality? This requires effort, will, discipline, imagination, and vision, at the very least." (Wiseman, 2016, p. 58)

Artificial moral enhancement

All of these issues have culminated in the creation of the most recent, third alternative for the project of moral enhancement – *the artificial moral enhancement*. This proposal aims to be more efficient and less controversial than moral bioenhancement as it keeps the benefits without the drawbacks, which worry enhancement critics (Giubilini & Savulescu 2018; Savulescu & Maslen, 2015). The main element of the suggestion entails the creation of artificial moral intelligence (AI).

[8] „...Of nineteen traits... respondents were least willing to enhance the most morally significant traits on the list: empathy and kindness." (Douglas, 2013).

The first version of the proposed AI system (Savulescu & Maslen, 2015) entails these capacities. First, it gathers and computes reliable data to alert the human agent on potential moral influences and biases. Second, it suggests strategies for ameliorating these influences and biases. Third, it advises the human agent on particular courses of action (Savulescu & Maslen, 2015, p. 84). In this regard, it serves as a moral monitor and a moral prompter. In doing so, moral AI aims to preserve the pluralism of moral values and to enhance the agent's autonomy. This is achieved by prompting moral reflection and by helping the human agent overcome her natural psychological limitations. So, prompting the user would be enough for „the agent to reflect on and assent to moral values and principles" (Savulescu & Maslen, 2015, 80).

Dobrocevsky has recently responded to this idea, arguing that it has some significant problems. First, moral AI is „being too tied up with the values of its users and not being able to provide a solution to the Moral Lag Problem" **(**Klincewicz, 2016). In being too dependent on the user's values, i.e., lacking in moral autonomy, the system has no power in changing the agents' attitude to overcome individual biases. This is especially the case if the human has either already „overcome these biases or does not feel that overcoming them is a part of their moral values..." And since the AI has no other options but these, its „advice becomes altogether useless....". **(**Klincewicz, 2016) Conclusively, this puts the effectiveness of the system in doubt as one requires a normative moral AI, not merely an advisory one, to tackle the issue of the Moral Lag seriously:

> „Being prompted to reflect on information from the environment or moral values and principles one endorses is not sufficient to lead one to act morally in situations where doing the morally right thing conflicts with these values." **(**Klincewicz, 2016)

Although not citing critics, an updated proposal (Giubilini & Savulescu, 2018) answers these critical points. It promises both greater efficiency and less controversy than prior biomedical approaches, which have usually worried enhancement critics (Giubilini & Savulescu, 2018). The authors name their proposal: *artificial moral enhancement* and, aptly, engage two vital points: its fine-tuned effectiveness and the preservation of human autonomy.

Concerning its fine-tuned efficiency, the moral advisor claims to produce person-specific, fine-tuned, expert-like advice. This kind of AI-generated advice is "more informed and more capable of information processing than any other human moral expert we trust" (Giubilini & Savulescu 2018, p. 10). Such personally tailored expert-like assistance fosters a type of moral self-evaluation for the human agent not available to other traditional or biomedical methods. It is also capable of engaging both dimensions of morality - reason and emotions. (Helion & Ochsner, 2018).

For instance, in the process of moral deliberation, authors propose that the AI system enhances moral decision-making by helping achieve a wide and narrow reflective equilibrium in providing counter-intuitive responses. As such, the artificial system allows individuals to develop their moral perspective within some fundamental moral constraints (for instance those derived from the right to life, personal freedom, and integrity), which the AI considers. Concerning emotions, the AI advisor aims to activate the right emotion for a given context to promote moral attitudes and moral action. By doing so, the human agents can become aware of distorting emotional factors that negatively affect their moral decisions and actions. Thus, it aims to implement the positive functions of intuitions and emotions without their downsides (Giubilini & Savulescu, 2018). Also, by making humans precisely aware of their emotional distorting factors, moral AI is capable of vastly improving moral actions. The reason lies in the psychological fact of how "intuitions and

emotions drive most of our moral and practical decision-making" (Giubilini & Savulescu, 2018, p. 16).

The proposal is also serious about both preserving and enhancing moral autonomy as it allows individuals to implement their moral perspective (in the best possible way) within specific necessary moral constraints that the AI considers. In doing so, the AI system preserves the pluralism of moral values. It also enhances the agent's moral autonomy by prompting reflection and by helping him overcome his natural psychological limitations (Giubilini & Savulescu, 2018). Still, the authors do not invest heavily in the question of what precisely these constraints ought to entail. They do settle on the conclusion of how they should include reasonable moral boundaries, already present in democratic societies. Finally, the proposal is also, starkly, cognizant of the unwillingness issue:

> "the AI would not be able to make us more moral unless we already have at least a basic willful commitment to be moral - which unfortunately many people do not have" (Giubilini & Savulescu, 2018, p. 181).

Although not giving a solution on how to fix this, the authors do propose to program the AMA with "basic filters" provided by moral experts. These basic filters would constrain the range of possible user given operational criteria to be used as the primary input for the AI system. As the authors note, the reason for instilling such basic constraints is not only theoretical but also very much practical.

> "We need to put constraints on people's behavior regardless of whether we think there are objective standards of ethics because there are pragmatic or political ends—most notably

> regulating the cohabitation of different people—whose justification need not be based on metaethical theories. Western society is already organized around some basic principles—such as reciprocal respect, tolerance, protection of persons' lives—which are enforced regardless of whether some people or moral systems acknowledge them. The AMA would simply follow the same approach we already adopt in shaping the institutions of liberal societies." (Giubilini & Savulescu. p. 171).

These ethical rules could then be utilized to prohibit the AI from condoning immoral behavior which would outline the space for valid moral exploration. As such, the human agents using the moral AI to pursue different moral goals would do so within reasonable moral boundaries, which promotes a pluralism of ethical views.

Unfortunately, the proposal does not delve further into the necessary moral constraints to ask the question of who decides what these moral constraints will be, and how do we choose the best among them?[9] This is of particular concern when we think about the wide distribution of such AI advisors, especially inside sensitive groups and children. Also, as the proposal does not engage the issue of children's enhancement, it cannot question if the introduction of moral AI assistance in early developmental phases improves fine-tuned effectiveness. It also does not engage the issue of unwillingness and lower rates of acceptance, which this kind of enhancement can quickly tumble on. Finally, the proposal does not connect with the broader social use of AI systems nor the

[9] This is the question of normative AI and machine ethics. Here, as we will soon see, the artificial agent which is not only an informer or advisor but an autonomous agent with the capacity for ethical reasoning. In other words, the artificial agent is capable of not only giving an answer about what to do but also, at least in some degree, why to do it.

research on AI ethics, both crucially important if we aim to accomplish a wide-spread distribution of moral enhancement.

To engage these issues, namely: fine-tuned effectiveness, unwillingness, and AI ethics this work provides the concept of symbiotic moral augmentation. I will first introduce the concept of moral augmentation. Here I point to the difference between artificial moral enhancement and moral augmentation and bring the first outline of the symbiotic relationship, which I will later fully develop as the socially acceptable and ethically valid means towards moral augmentation. Let us begin.

1. MORAL AUGMENTATION

> "Some people call this artificial intelligence, but the reality is this technology will enhance us. So instead of artificial intelligence, I think we'll augment our intelligence." Ginni Rometty

The change of the concept from moral enhancement to moral augmentation is deliberate. I take that moral augmentation rather than artificial moral enhancement is more fitting for the possibility of improving human morality through ethical AI systems. I offer two main reasons for such a change—a philosophical and technological reason. With the first, I aim to distance my proposal from some underlying philosophical assumptions present in the enhancement concept. With the second, I aim to base my proposal on an optimistic but sober technological realism innately connected with the idea of human-machine partnership. I aim to elucidate, that both of these are interconnected and, in the case of augmentation, naturally lead to the prospect of symbiotic human-AI collaboration.

1.1. Enhancements and posthumanism

The first reason, one of a direct philosophical outline, lies in the incentive to distance this proposal from the underlying moral impetus of the enhancement imperative (enhancement or bust! [10]) and some enhancement related tenets of transhumanist philosophy – inside which the project of enhancement was initially developed and with which it often stands related. By doing so, I hope to provide a more pruned, and acceptable proposal for moral improvement which naturally aligns with the expected technical development and social implementation of AI technologies in the close

[10] Wiseman complains that the impetus of morally enhance or die has „really made a joke of this domain“and he hopes that this approach may „be abandoned by commentators completely, leaving nothing over and that it never be spoken of again. “ (Wiseman, 2016, p. 263).

future. It is, thus, essential to present some crucial points on transhumanism. Transhumanism understands itself as a life philosophy, an intellectual and cultural movement, and an area of study.

Ever since its earliest contemporary conceptions, (More, 1990) transhumanism understood itself to be rooted" in secular[11] humanist thinking" (Bostrom, 2005, p. 4) [12] and has two main goals – enhancement and posthumanism. The first goal, that of enhancement, entails fundamentally improving "the human condition through applied reason, especially by developing and making widely available technologies to eliminate aging and to greatly enhance human intellectual, physical, and psychological capacities." (Transhumanist FAQ - Humanity+, 2020)

The reason why we should engage in cognitive and physical enhancements lies in the severe existential risks of this century. They include not only food and water shortages, global warming, economic instability, wars, but also the catastrophes resulting from technological developments (Bostrom, 2009).[13] To ensure our survival, the argument goes, we need to save our species from the vexing existential threats and possible destruction, for our evolved biology is simply not fit enough for the dangers and challenges that await us. This entails redesigning our current human condition, including parameters such as the inevitability of aging, limitations on human and

[11] As Max More: "Transhumanism could be described by the term "eupraxsophy" coined by secular humanist Paul Kurtz, as a type of nonreligious philosophy of life that rejects faith, worship, and the supernatural, instead emphasizing a meaningful and ethical approach to living informed by reason, science, progress, and the value of existence in our current life." (More, 2013, p.4.)

[12] Here it is important to note that the secularist dimensions of transhumanism have, through the years, experienced a mellowing which I perceive as a natural development of the theme's wider academic adoption and the engagement of religious scholarship with transhumanist thought. (Mercer & Trothen, 2014).

[13] It is interesting to note here that Bostrom follows upon this kind of "existential risk" narrative even in his most recent artificial intelligence based work. Moreover, his approach to the theme of artificial intelligence is exactly one of existential danger. In doing so, his approach has received contention from many AI experts, as we will see a bit later.

artificial intellects, suffering, and lastly our confinement to the planet earth (Bostrom et al., 2009). In short, we need to take charge of the evolutionary[14] process through technological expertise to liberate the human species from its biological limitations (Young, 2009). In the initial enhancement discussions, this predominantly included cognitive and physical limitations (Transhumanist FAQ - Humanity+, 2020). However, the subject shortly moved to include the enhancement of human morality even as the motivation of existential risk remained unchanged.[15] This was the introduction of the original moral enhancement position (Persson & Savulescu, 2008) and the concept of "moral transhumanism" (Persson & Savulescu, 2010) which argued that,

> "if human civilization is to avoid destruction or deterioration, human beings need to become more human in the moral sense. Such morally enhanced humans may be called transhumans or posthumans." (Persson & Savulescu, 2010, p. 13)

Here then, already in the original moral enhancement reflections, we have both transhumanist goals, which include enhancements, as means by which we surpass our inherent limitations, and posthumanism as the final state, the end goal, of the enhancement process. From the perspective of its final goal, transhumanism is then perceived only as an intermediate state, one in which humans find themselves as they apply enhancements, eschew the existential risks, and aim to achieve the posthuman state. To achieve the posthuman state one then needs to traverse a series

[14] Kahane and Savulescu (2015) summarize this point when they state how "the current levels of" naturally existing substances and processes "are set by largely blind processes, meaning that it's highly implausible that they are already at an optimal level, relative to our values". (Kahane & Savulescu, 2015).

[15] "It might be reasonably doubted that there is enough time for human beings to undergo the requisite degree of moral enhancement before it is too late, before they put their formidable technological powers to catastrophic use. " (Perrson & Savulescu, 2012, p. 2).

of enhancements in body and mind which now include the full package of cognitive, physical, and moral enhancements. As More explains, we can

> "thoughtfully, carefully, and yet boldly apply technology to ourselves, we can become something no longer accurately described as human – we can become posthuman". (More, 2013, p.4)

In obtaining the posthuman state humans are envisioned to fully exceed the limitations which define the less desirable aspects of our current human condition.[16] This includes not only having much higher cognitive capabilities and more refined emotions (Transhumanist FAQ - Humanity+, 2020) but also being rid of the disease, aging, physical suffering, and death. This is also the crown achievement of posthumanism, a state of (digital) immortality, a novel existence made possible for us through the release from the frailty of our current materiality. To achieve this state of surpassing bliss and a higher state of being[17] (Bostrom, 2010), we need to create for ourselves a new form. This form allows for greater freedom and ability to achieve knowledge, experience, and in the end - transcendence. Thus, to achieve posthumanism, it is not enough to modify the biology, we have to release ourselves (our minds) from it. Posthumans are then imagined to have not only vastly

[16] This is exemplified in the Why I Want to be a Posthuman When I Grow Up (Bostrom, 2013). Here Nick Bostrom envisions how the posthuman existence possesses capacities (healthspan, cognition, or emotion) significantly exceeding the maximum attainable by any current human being. Bostrom argues that some possible posthuman modes of being would be excellent and that it could be excellent for us to become posthuman.

[17] As Bostrom, adopting a narrative posthuman character addressing contemporary humans writes:" You could say I am happy, that I feel good. That I feel surpassing bliss and delight. Yes, but these are words to describe human experience. They are like arrows shot at the moon. What I feel is as far beyond feelings as what I think is beyond thoughts. Oh, I wish I could show you what I have in mind! If I could but share one second with you!" (Bostrom, 2008).

greater capabilities but also freedom of form – often referred to as morphological freedom (More, 1993; Sandberg, 2013). As Hauskeller noted,

> "The telos, the logical endpoint, of the ongoing cyborgization of the human is thus the attainment of digital immortality." (Hauskeller, 2012)

To conclude, the transhumanist philosophy and original moral enhancement proposals take that the existential risks, made possible with our technological progress, gravely jeopardize humanity's long-term survival and beneficial development on our planet. To ensure our survival a wide array of enhancements, cognitive, physical, and moral is proposed. Lastly, the utilization of enhancements not only ensures our survival but also, through a continual application, allows humans to become posthumans - to eschew death itself.

1.2. Sober transhumanism

How does my proposal diverge from this original notion of moral enhancement? The difference lies in both of its two important dimensions. The first point of difference is posthumanism, as the final goal of enhancements, and the second point is the understanding of enhancements as being intensive (and often intrinsic) in their application and urgent in motivation – an interpretation alleged by the existential risks humanity faces in the 21st century. As previously noted, these two are often correlated since the only way to fully release humanity from the ultimate kinds of existential threats (disease and death) assumed by authors is to achieve the posthumanist state.

First the final goal of enhancements. For transhumanism, this is posthumanism. For my proposal, the final goal is far more modest – moral augmentation in everyday contextual scenarios. Does my proposal, then, exclude the possibility of posthumanism? Not necessarily. However, it also doesn't

include it as the final goal of augmentations. This means that I do not negate the possibility of achieving some kind of a posthumanist state (in the far, far future) nor do I negate that symbiotic artificial intelligence might be one of the means to get us there. I simply do not relate the two as I wish to disconnect the augmentation proposal from the, often, ostentatious posthumanist ideal. I take that this move improves the credibility and practicality of the proposal. Especially if some of the more contentious transhumanist ideas such as morphological freedom or mind digitalization have been evaded by substituting the posthumanist goal with a more modest and practical one. Additionally, this also releases the proposal from justifying why the goal of enhancements naturally leads, or ought to lead, to the posthuman state, and why this state is even a worthy goal to achieve.

However, here I find it prudent to notify that the motivation for escaping these notions, does not lie in the transhumanist's materialistic and often secular worldview. My proposal has no quarrels with materialism nor secularism. Rather, I propose a full acceptance of methodological naturalism and moral reasoning which refrains from religious explanations, as means by which we achieve answers about human morality which can be then technologically utilized to improve that same morality. In this sense, my proposal posits itself within the framework of philosophical and ethical engineering (Palermos, 2019) rather than philosophical or scientific speculation as it aims to follow a simple rule, good science, good philosophy.

This also brings me to the second point – that of enhancements. In the originally outlined proposal of moral enhancement, enhancements aim to intrinsically impact the biological and psychological limitations of individuals. The approach aims to justify the urgency of the application by the allegedly looming existential risks. I say allegedly as I do not share this concern on which I exemplify further on.

Summarized, and in order of the argument, the alleged existential threats urge the application of moral enhancements, enhancements are focused on the individual, enhancements aim to intrinsically alter the individual's biology and psychology. [18]

My proposal differs on all three points. First, it does not derive its motivation from the alleged existential threats of our current planetary condition. The reason behind this position is that moral augmentation of humanity through AI means is not directly related to existential threats but can be easily accepted as the logical outcome of developing symbiotic and ethical artificial intelligence. The justification behind this conclusion, as we will later see in detail, entails that both humans and machines are fallible and that we necessarily require each other's strengths to improve each other's conditions. In other words, if we aim to pursue beneficial artificial intelligence applications, then the symbiotic partnership is entailed as the ethical outcome of such an endeavor.

Importantly, this move also disentangles the project of moral augmentation from the urgent imperative (Persson & Savulescu, 2008) to morally enhance humanity - as that dance hit goes -" right here, right now". [19] This is the second point of difference. Here, painting a grave picture of

[18] (Bublitz and Paulo) have recently summarized these points, based on the Perrson Savulescu's comprehensive proposal. "1. To avert ultimate harm to humankind, we urgently have to address, among other things, the problems of climate change, global injustice, and dangers from weapons of mass destructions. 2. These problems are the result of morally defective behavior. 3. Morally defective behavior is caused by psychological deficits in individual minds, e.g., temporal and spatial parochialism. 4. Solving these problems requires remedying these individual-psychological defects. 5. Traditional means of enhancing moral behavior have proven incapable of redressing these problems, possibly because our genetic or biological make-up has emerged in conditions very unlike those of today. 6. In the not too distant future, biomedical means might be developed that afford the altering of those biological and psychological deficits. 7. There is no morally relevant difference between biomedical and traditional means of improving moral behavior (parity principle). 8. Therefore, we urgently need to develop and use biomedical means to improve moral behavior." (Paulo & Bublitz, 2019, p. 97)

[19] Fatboy Slim. "Right here, right now". (Slim, 1999)

existential threats and positing a strong urge to fix them is not required to justify nor to motivate the prospect of moral augmentation. This is especially important to note since the augmentation proposal does not include revolutionary modifications to human biology or psychology as do some of the biomedical enhancement proposals. This is the third point of difference. Here the proposal evades the strict focus of "methodological individualism" [20] which focuses on the intrinsic modification of one's biology (Paulo & Bublitz, 2019). Rather, the augmentation proposal focuses on achieving moral improvement through a symbiotic partnership. This entails augmenting (Kahneman et al., 2016; Kleinberg et al., 2018) oneself through collaboration (Jennings, 2018) with the AI partner's capacities rather than substituting or intrinsically altering the individual human's capacities. This approach allows for two beneficial outcomes.

First, it allows moral augmentation to build upon human capacities as these can be supplemented and augmented by AI systems, rather than be altered or deleted, by intrinsic biological modifications. This allows the project of moral augmentation to avoid the harm of producing unwanted biological modifications. These also include the more problematic, and inherent, dangers of persistent and undesirable changes in our moral psychology as we aim to enhance a specific limitation.[21] For instance, if we would aim to modify "counter-moral" emotions, such as

[20] "The view that all higher level social processes can ultimately be exhaustively explained (and, hence, remedied) at the level of the individual. It presupposes the reducibility of societal, cultural, historical or economic matters to the behavior of individual persons (and their psychology or even neurobiology)." (Paulo & Bublitz, 2019, p. 100)

[21]The biggest problem I perceive here is that what I call the "Eloi conundrum" – where morally advanced humans become incapable of surviving in the harsh world of unenhanced ("Morlocs"). Alternatively, we could forcefully change the current emotional make up, (for instance by suppressing our emotions) and devoid ourselves from the capacity to morally assess a perfectly ordered but immoral society. Here, again, science fiction has provided us with an example in the movie Equilibrium which deals with biomedical technologies. However, a more realistic scenario entails utilizing

reactive, heavy, aggression, through biomedical means we must have in mind, as Wiseman (2016) poignantly notes, that

> "removing a person's aggressive impulses is only going to result in an overall net benefit of moral goods in circumstances where that aggression is being applied in comprehensively morally problematic ways. The fact is that aggression informs a huge number of various activities we perform on a daily basis, many of which are morally neutral (e.g., sports), and indeed in certain circumstances, aggression can be a necessary evil, and in extremis a righteous moral good. So, an impulse to do a good thing may well actually benefit by being enriched with a certain amount of aggression, again, such as intervening to prevent a hostile individual doing a tremendous harm." (Wiseman, 2016, p. 49).

Second, by evading a strict focus on the individual the moral augmentation proposal easily includes "social and environmental factors" (Paulo & Bublitz, 2019) which are often required to resolve the complexity of our existential moral issues and achieve wide moral progress. (Buchanan & Powell, 2018; Pinker, 2012). This means to say that the utilization of symbiotic artificial intelligence in producing moral augmentation with human beings, both individuals and societies as a whole, is a moral project that does not intrinsically jeopardize the most humane social and political achievements of liberal democracies reflected in the beneficial work of our social institutions.[22]

AI means (by governments or corporations) to „scan" one's affective states to evaluate a pre-determined allowable public set of behaviors. (I.e. we want our workers to be happy, not sad!).

[22]For instance, as Lilley (2013) remarks, both Jurgen Habermas and Francis Fukuyama have famously criticized the transhumanist project as being capable of undermining the value of the liberal state. Habermas asserts that..."if transhumans, and then posthumans, become cognitively or emotionally distinct from humans, commonality will cease and the polis will fracture. Habermas also warns that genetically-designed children will not be perceived as

Moreover, it counts on these for its success as it aims to improve our morality through ever-rising moral standards, achieved by the assistance of our AI partners. As such, the application of moral AIs is not an urgent and quick-fix method but is rather a gradual, safe, and systematic change that counts on cultural and social innovation. (Buchanan & Powell, 2018)

However, this does not entail that the augmentation proposal is intrinsically less effective than the possible achievements of the enhancement approach – an assumption that might be read from its initial humble goals. Just the opposite, symbiotic partnership as means of moral enhancement, could be, at least partially, tested within the next decade[23] without provoking any grave harm to human dignity or the stability of the societal order. Moreover, if we want to, the symbiotic partnership can also be utilized to conceptualize the achievement of the transhumanist dream (as so many science-fiction scenarios have done) – it is merely not necessary to do so. In this sense, the concept of moral augmentation is philosophically modest, but also quite optimistic about its achievements – it promotes itself as an optimistic technological realism.

1.2. Optimistic technological realism

By proposing an optimistic technological realism I mainly aim to affirm two crucial points. First, I focus on realistic rather than outlandish AI technologies, and second, I focus on augmentation rather than substitution of human agency. By following this route, I hope to seamlessly arrive at the concept of symbiotic relationship.

---

autonomous actors because they will be governed by the irreversible intentions of third parties. Once again, at stake is universal egalitarianism." (Lilley, 2013, p. 8.)

[23] If I have to position a probable date, I would put the date closer the end or the beginning of the next decade.

With the first point, I eschew the kind of opportunistic futurism, which was often present in enhancement debates. Also, since AI research is, on its own, quite susceptible to hype, it seems unreasonable to inflate these two research fields together. Following, I will not utilize the possibility of artificial general intelligence (AGI), and neither will I discuss super-intelligence. Additionally, I do not wish to discuss how much of existential risk (in terms of superintelligence) does AI pose for humanity. I do acknowledge the philosophical and social importance of researching such themes, and I utilize some valuable insights generated there. Still, I take how these debates belong to a more theoretical domain which is not related to the close and medium-future of AI research. By doing so, I also wish to refrain from discussing the possibility of developing artificial moral agents, or full moral agents. These are artificial agents with some kind of moral consciousness and subjectivity, thus moral persons and moral subjects.

To restrain from venturing into such inspiring, but for the hypothesis rather profitless, debates I shall follow that what, Andrew NG, calls the "virtuous AI circle" (Ng, 2017). Here, the aim is to engineer safe, quality products capable of generating trustworthiness, augmenting human capacities and operations, and advancing both AI and moral research. And they are capable of such fundamental openness because they are practical instantiations, artificial systems co-existing with humans in the shared techno-social world. The opposite is, as NG calls it, the unvirtuous circle of hype, or "evil-AI" hype. This attitude brings attention to concerns about "evil-AI" technology which deals with the fear of AI-induced existential threats (Karnofsky, 2016), such as those produced by advanced AI, for instance where the super-human AI becomes a complete master over humanity's destiny (Bostrom, 2016). The fear of these AI-driven global catastrophes (Turchin, 2020), then, shifts focus (and funds) from more practical, impactful, and directly observable problems. These include the issue of automation (Ford, 2015) or human-AI collaboration, the problem of AI bias or

explainability, and the issue of epistemic and moral de-skilling. Finally, the asymmetry of power, which is a further exacerbation of political and economic power through AI, unfortunately already manifesting in the use of social networks. (Pew Research Center, 2020). By staying within the virtuous, realistic AI circle, I also change the focus from a more theoretical one (super-intelligence and AI-safety) towards a more implementable one (machine and roboethics).

This entails that I will focus on artificial agents as functional moral agents[24]. As Moor's influential 2006 paper articulates, here I am thinking of the explicit ethical agent, which is "explicitly" guided by a normative ethical model while operating in a particular ethical scenario (Moor, 2006). For instance, a software agent that safely drives the autonomous vehicle from point A to point B, and is capable of providing valid ethical responses to different ethical dilemma's on the road would be one such explicit agent. Unfortunately, we are still far from achieving fully functional ethical reasoning in the public realm as the complexity of public-life often surpasses the possibilities of our current ethical models. However, this is a realistic goal we are working towards. In the meantime, we are dealing mostly with the other two types of ethical machine agency – ethical impact agents and implicit ethical agents. The ethical impact agent includes any kind of software or robotic agent which can be ethically evaluated for its actions. For instance, a traffic-light safeguards human well-being in the public space and with its actions creates ethical outcomes

[24] Wallach and Allen distinguish operational, functional and full morality. Machines with operational morality are instilled with simple or complex ethical rules but they are not much different from an "ethical matrix". There is no autonomy present. Functional morality on the other hand already marks a huge step forward in the autonomy of the moral decision the machine makes since it can create moral decisions without them being directly caused (blindly follow) the moral instructions given by its designers. Machines on this level cannot be predicted in their moral autonomy. Finally, the third level is that of the full morality which would entail an artificial agent endowed with full moral autonomy and responsibility. It is presumed this could not be made possible without a form of self-awareness. (Wallach & Allen, 2008)

even though it is not instilled with specific ethical programming. If the agent is instilled with ethical programming but has no capacity for ethical reasoning but merely ethically reacts, rather than acts – then we are dealing with an implicit ethical agent. Here, for instance, we can think of a vacuum-robot being instilled with safety (ethical) measures that prevent it from spilling a glass of hot water over the relaxing cat or a child that crawls on the floor. The main take-away of Moor's (Moor, 2006) and Wallach and Allen's (Wallach & Allen, 2008) evaluation I aim to utilize here is to escape the all vs. nothing attitude which states that either we have machines with moral capacities similar to humans (and a corresponding moral status), or we do not have moral machines at all.

Hopefully, this will bring about the conclusion of how even though our machines may be fundamentally ethically limited, they are capable of achieving human moral augmentation – as interactive ethical means. Moreover, inside a symbiotic relationship, they form human-AI partnerships[25] that outperform the capacities of both individual humans or individual machines. Furthermore, the importance of "symbiotic" within the human-machine relationship entails that it is not only the human benefiting from the partnership but rather the machine too, and the overall, joint, human-machine partnership. This is the second point I wish to uphold – augmentation.

Augmentation is crucial since already, in our present time, it is clear how AI systems are outperforming humans in many tasks of both physical and cognitive agency. However, as machines are gradually becoming super-performers, they are not becoming generally more intelligent. This fact already showcases how even though AI systems may never become super-intelligent, they are capable of becoming highly specialized super-agents. Moreover, this kind of super-agency is not

[25] The idea of machine partners, rather than machine servants or machine overlords is getting important traction not only for civilian (Markoff, 2016) buth military AI research as well (*Drubin*, 2019).

limited only to substituting human operations but also in supporting and empowering them. In other words, artificial systems are not only capable of automating but also augmenting human agency.

This makes augmentation rather than automation, or substitution of human agency, the dominant issue for AI research in the 21st century. The reason for such a strong claim is simple. The nature of artificial technology is similar to electricity, it is here to stay with us for a long time, and it is poised to transform our livelihood. This livelihood includes not only the industry, the meaning, level, and scope of social work, but also the crucial elements of human existence - our leisure and daily lives, our social and personal relations, our decision-making, and understanding of ourselves.

Moreover, since every part of the human agency, will get gradually impacted by the capillary distribution of AI systems, a dominant scientific, political, and philosophical question is raised. What kind of cooperation can we achieve with machines so that they preserve and augment rather than diminish or squander human potential?

1.3. Symbiosis and augmentation

Some of the most severe attempts of exploring augmentative human-AI cooperation are found in recent research proposals, such as the human-centered AI, trustworthy AI, shared autonomy systems (Fridman, 2018), active-learning systems, and finally, symbiotic systems (Raisamo et al., 2019). What is crucial to note here is that all of these approaches are fundamentally incited by some of the necessary technical limitations the current and close-future level of AI technology exhibits. These include not only the technical inability of the machine to reliably and continuously operate in a fully autonomous manner but also the plethora of severe ethical harms the use of AI systems might and does produce. For this reason, serious AI proposals are aware of machine

limitations, withhold value and ethics as their common denominators, and aim to develop AI systems that have the power to augment human agency rather than substitute it.

And in doing so, they are inadvertently continuing upon the vision of early cybernetics scholars such as Licklider (1960), who was the first to envision symbiotic human-AI systems, or Douglas Engelbart, his contemporary, to whom the concept of augmentation is initially credited. He, for instance, wrote how augmentation is a process by which machines are:

> „Increasing the capability of a man to approach a complex problem situation, to gain comprehension to suit his particular needs, and to derive solutions to problems". (Engelbart, 1962)

What is crucial to note here is that the first augmentation theorists placed the process of improving humans already within the symbiotic framework. Here, the machines and humans are partners, endowed with autonomy, and the capacity for interactive collaboration. As such, they can form a single system greater than the sum of its parts. This crucial distinction also points us to answer a vital augmentation question: Why should we choose symbiotic systems rather than some other means to (moral) augmentation?

To answer this question, I must first recall how the AI system enhances the human agent's moral decision-making by helping achieve a wide and narrow reflective equilibrium (Giubiliini & Savulescu, 2018). The method of reflective equilibrium[26] is a deliberative process, a method, or a way of

[26] The method has been officially defined by John Bordley Rawls in his Theory of Justice (Rawls 1971) as means by which we can define a a set of coherent and generally applicable principles that will establish a stable social order. The general principles we obtain and refine by the method of reflective equilibrium will be those which are rooted in our sense of justice and thus withhold inherent motivation to follow them. By utilizing the reflective equilbirum Rawls obtained the "original position".

thinking by which we bring into coherence, systemize or order, our moral judgments (convictions) about a particular moral case with ethical principles we accept as being generally applicable. As Rawls, initially describes:

> "By going back and forth, sometimes altering the conditions of the contractual circumstances, at others withdrawing our judgments and conforming them to principle, I assume that eventually, we shall find a description of the initial situation that both expresses reasonable conditions and yields principles which match our considered judgments duly pruned and adjusted. This state of affairs I refer to as reflective equilibrium. It is an equilibrium because at last our principles and judgments coincide; and it is reflective since we know to what principles our judgments conform and the premises of their derivation." (Rawls, 1971, p. 18).

Here, then, if the system expresses a piece of advice on a specific moral choice we deliberate, it can aim to "balance our intuitions against the piece of advice, and vice versa, to attain a condition of reflective equilibrium." (Giubilini & Savulescu, 2018, p. 12). However, to do so, the system has to have autonomy over its actions, making them different from those of the human agent. This makes the AI system an agent-system or an artificial agent. As Russel and Norvig explain:

> "An agent is anything that can be viewed as perceiving its environment through sensors and acting upon that environment through actuators…A human agent has eyes, ears, and other organs for sensors and hands, legs, vocal tract, and so on for actuators. A robotic agent might have cameras and infrared range finders for sensors and various motors for actuators. A software agent receives keystrokes, file contents, and network packets as sensory inputs

and acts on the environment by displaying on the screen, writing files, and sending network packets." (Russell & Norvig, 2002, p. 34).

"Of course, all computer programs do something, but computer agents are expected to do more: operate autonomously, perceive their environment, persist over a prolonged time period, adapt to change, and create and pursue goals… A rational agent is one that acts so as to achieve the best outcome or when there is uncertainty, the best-expected outcome". (Russell & Norvig, 2002, p. 4).

Moreover, since the AI agent has autonomy, the type of system agency which is constituted by the artificial agent and the human agent is not one of an atomistic, individual, agency where only the human has autonomous agency. It is also not a type of extended agency where the artificial system has no autonomy (Hernández-Orallo & Vold, 2019). Instead, it is a type of distributed or joint agency where both involved agents, the human and the machine, are autonomous, even though their autonomy is not of the same kind or level.

The main difference between the distributed and extended types of agency, to exemplify shortly, is that in distributed or joint agency, there are minimally two or more autonomous agents among whom the goals and goal-obtaining actions are distributed (according to a specific pattern). In systems of extended agency, on the other hand, there is only one autonomous agent, ordinarily that of the human, and the non-autonomous system which „extends" the capacities of the human.

Thus, in the distributed agency, including those of human-AI multi-agent systems (Pacherie, 2013), the locus of the agency is on the compound unit of the system, the group, and not on the individual. Here the agents intent as one, act as one, and are also normatively evaluated as one. Similarly to

the musketeer's slogan:“ all for one and one for all“ in distributed systems the constitutive agents are not individual but collective intention and action - the content of their intentions is always the collective “We” and not the individual “I.” (Pacherie, 2013) As such when an individual agent plans to act out on the shared goal, the content of her intents is always in terms of “we are doing this” or “I intend that we do something.” This point also showcases how the subject, the bearer, of such collective intentions is always the individual agent and not some emergent collective mind which subsumes, gobbles up, individual agency, and autonomy. In other words, the individual agent never loses itself inside the distributed agency. Instead, the individual agents' attitudes are recognized and valued as it is precise that in the sharing of their attitudes the complex (shared) intentions can be formed and acted out. (Schweikard, 2017).

Extended systems, on the other hand, do not include more than one type of autonomy – usually instantiated within a single kind of an autonomous agent – for instance, a human. Instead, extended systems are constituted from autonomous agents and “extenders” (Hernández-Orallo & Vold, 2019). These can include systems, tools, artifacts, or processes which extend the (cognitive and physical) operations of the autonomous agent above and beyond her capacities’ range.

In the case of AI, systems extenders can practically entail different AI services. For instance, augmented reality systems, language translators, or the recent iteration of writing assistants (such as Grammarly, with which I am augmenting the here written text) can all be understood as AI extenders. Even though they have no autonomy of their own, they are supporting, empowering, and extending the already existing capacities of the human agent. Crucially, they cannot function

without the human agent on whose autonomy their service is based. [27] Poetically, extenders operate only when they are connected to and utilized by a human agent. Their autonomy is human autonomy, although their capacity for agency is not the human's capacity for agency.

For instance, as exemplified in Merleau-Ponty's blind man's cane example, the cane (through constant use) can become not only an object but also a vehicle of his awareness. When this happens, conscious interpretation or control is no longer needed, and expanded perceptual and experiential capacities are established. (Schweikard, 2017). As such, even though the AI systems are capable of enhancing human capacities, by extension[28], they are not capable of forming collaborative, interactive, partnerships as autonomous agents:

> "Note that the lack of autonomy of *E* is crucial to see this as an Extension rather than a collaboration. This detachment between cognition (or even intelligence) and autonomy is well-aligned with the view of cognition as a service, where several facilities for visual perception, speech, and language processing are provided, as well as other inference and

[27] For instance, Grammarly cannot write anything autonomously for me. All the decisive input of this text comes from me, Tom, the author. I had to write it down, for it to exist. But, once I have written my sentences down, the Grammarly assistant has provided many corrections and improvements to the written text. In this regard, this text is truly the product of augmented Tom: Tom, augmented by extension of Grammarly. However, if Grammarly was capable of autonomous writing, and could complete entire sentences on my behalf (or even paragraphs), then I would stand augmented by collaboration rather than "mere" extension.

[28] In its most recent interpretation (the so called second and third wave) extended cognition explores an interdisciplinary study of the variety of relationships and different kinds of interactions that can occur between human agents and cognitive artifacts. It researches the ways human cognition is constituted in the world and the ways through which cognition can work in conjunction with technological artefacts and tools (Menary, 2010, p. 227). "Complementarity accounts stress the different contributions made by internal and externally located items, arguing that it is precisely these differences that allow for the environment to play a transformative role in cognition". (Kiverstein & Farina, 2011, p. 3).

> reasoning solutions, independent of any Task. For instance, an online translation system can be provided as a service, which can be integrated into many kinds of applications and goals". (Hernández-Orallo & Vold, 2019, p. 510)

Human-AI symbiosis, on the other hand, focuses on the synergetic outcome produced by the dynamic and profoundly integrative collaboration of autonomous agents. But as previously noted, to provide effective augmentation, especially moral augmentation, the AI agent has to be personally tailored, that is „fine-tuned" to its human user. This entails that the agent not only has to be capable of providing concrete answers for specific moral scenarios, but it is also personally tuned to the user's character. It has to be in tune with the preferences, beliefs, motivations, and other fundamental character traits of their users. This kind of fine-tuning to the everyday cognitive and emotional processes of their users allows for precise augmentation. So, the symbiotic agent has to exemplify a type of connection with its user, which is similar to the one depicted in extended systems.[29] As Licklider, so quaintly noted, more than thirty years before the paradigm of extended cognition:

[29] Heersmink has contributed extensively in this regard when as he develops a precise hermeneutic to assess the dimensions cognitive artefact should have in order to establish such close coupling. „These dimensions include the kind and intensity of information flow between agent and scaffold, the accessibility of the scaffold, the durability of the coupling between agent and scaffold, the amount of trust a user puts into the information the scaffold provides, the degree of transparency-in-use, the ease with which the information can be interpreted, the amount of personalization, and the amount of cognitive transformation. Generally, if the integration is dense, the artifact can be seen as part of an extended cognitive system, whereas if it is shallow, the artifact can be seen as part of an embedded cognitive system "(Heersmink 2017, p. 437).

> „To think in interaction with a computer in the same way that you think with a colleague whose competence supplements your own will require much tighter coupling between man and machine than is suggested by the example and than is possible today. “ (Licklider, 1960)

Furthermore, it is precisely this fine-tuning, this close coupling, towards their users, which builds familiarity and trustworthiness. This provides a more comfortable and seamless acceptance of the technology. Also, since the global society expects ethically and socially acceptable AI systems, it seems only natural to connect the research on ethical machines with the research on moral improvement through (ethical) machines. For instance, the ethical utilization of cooperative artificial assistants in (children) education is one of the two dominant interests in AI research (the other being healthcare). If society is already investing in the development of educational and tutoring agents, then society can also investigate how to utilize this technology to provide wide-spread moral improvement. Especially when there is nothing objectionable to the idea of how it is good to be morally better. This point, at least, can be safely agreed upon. (Pinker, 2011). Conclusively, by arguing for human-AI symbiosis, I also aim to showcase why the concept of technological, moral augmentation can be justified, provide another model of moral research, and indicate the limits on the implementation of symbiotic technology. I take that this kind of personalized fine-tuning, achieved if the artificial companion grows with their human user from an early age and is correspondingly ethical – can effectively tackle the issue of unwillingness.

However, to have a lucid understanding of the symbiotic relationship, I first have to explore the technical limitations and opportunities which the current and close-future level of AI technology exhibits. In doing so, I aim to clarify why these limitations do not only make AI ethics a requirement but also naturally lead to the development of symbiotic systems. Enter machine learning.

2. MACHINE LEARNING: WHAT MAKES THE MACHINES TICK?

"If you could train an AI to be a Buddhist, it would probably be pretty good". Reid Hoffman

2.1. Short history

During the summer of 1956, several scientists, including Marvin Minsky, John McCarthy, and other notable figures, started a workshop on the Dartmouth College campus intending to brainstorm on the possibility of intelligent machines development. The motivating idea behind the workshop was to brainstorm how the use of mathematical logic and its implementation within formal programming systems could lead us to the development of autonomous intelligent machines. Due to the unappealing relations, McCarthy had with the cybernetic movement, and the need to differentiate their initiative from the cybernetical idea, McCarthy, coined the term – Artificial Intelligence.[30] The basic idea was to symbolically describe expert knowledge of a specific domain and then to program the machine with that knowledge. Hopefully, the computer could then simulate this kind of intelligent behavior. The first such trial included the game of chess, as McCarthy himself was an avid chess player.

Unfortunately, the computational power to run the program severely lacked at that time, and the entire list of moves and counter-moves had to be hand-coded into the program, so the idea was soon scrapped. However, another scientist took the paradigm and posited a novel approach to the problem. He envisioned a situation where the computer instead of being programmed with rules,

[30] This title remains still with us, even though the techniques responsible for the contemporary breakthroughs in artificial intelligence systems are predominantly influenced not to the Dartmouth thinkers but rather the opposite camp: the cybernetic movement. These were inspired by a system-based design inspired by living organisms, and especially the simple but highly effective concept of the "feedback loop".

learns the rules on its own, automatically. This scientist was Arthur Samuel. Samuel used the game of checkers, as he collected data from about a thousand games of Checkers, and allowed his program to write its own rules of play by learning from this data. This case was the first proper example of machine learning (McCarthy & Feigenbaum, 1990). And although Arthur Samuel's program was not capable of learning at an expert level of Checkers, the machine learning approach he initialized, accomplished heights few could have expected. For instance, Google's AlphaGO clean 3-0 victory against Lee Sedol, arguably the best contemporary human Go player in the world in 2016, is hallmarked as one of such successes (Wang et al., 2016). AlphaGo was useful because it has learned on millions of expert moves and could predict each of the movement's success through machine learning, thus adjusting its decisions on the fly. However, the next version of the AlphaGo program, AlphaGo Zero, surpassed the strength of the Alpha Go (Lee) in mere three days by a stunning 100 to 0 games as it exceeded all other versions of the AlphaGo program. Moreover, it did so in only 40 days, entirely from self-play without any human intervention or by learning from records of other GO games (Silver, 2017). And how was this super-human capacity achieved? AlphaGO Zero's success predominantly through reinforcement learning and AlphaGoLee through supervised learning. However, before engaging these techniques we have to take a look at the fundamental constraints and advantages of machine learning procedures.

2.2. Structure and underlying mechanisms

One of the crucial abilities which allow machine learning algorithms to represent and learn from data is feature extraction. Feature extraction, as the title says, is the ability to extract crucial features from a specific data set. In the language of medieval philosophy, which was one the first to deal with the way humans form abstract knowledge, these crucial features were described as

"quidditas" or "what it's likeness." That what makes a specific thing what it is. By using a set of cardinal features of a particular entity, humans can accurately identify an entity and compare it with another, to find shared similarities or differences. Additionally, humans can represent those crucial features in different formats, a procedure any student of art is familiar with. And the more examples one learns on the more proficient she becomes in detecting and representing these crucial features (for instance masterfully playing Mozart on a Stradivari Violin takes decades of practice).

And exactly this ability to learn on numerous examples is the strength of machine learning, specifically deep learning neural networks. The reason behind this lies in the deep learning networks' ability to detect and extract more features from a higher amount of provided data. The network is then capable of easily surpassing human-levels of feature recognition and representation as the network can automatically encode not only thousands but millions of such features and their hierarchical relations. Furthermore, this ability allows neural networks to approximate (often with great success) any arbitrary mathematical function. And all of these operations are fundamentally based on the operations of a single artificial neuron. Let us, then, take a closer look at this simple yet so efficient structure.

The artificial neuron is the fundamental building block of the artificial network. It is a computational unit, designed to roughly approximate the work of a biological neuron through mathematical operations. The single artificial neuron is a simple computational unit, but once stacked together in the form of neural networks, artificial neurons are capable of producing complex computations. How does the artificial neuron work? For a more natural understanding, I will represent the artificial neuron in three vertical columns. The first column on the left represents the part of the

neuron where the input arrives, represented as a vector of numbers. The middle column represents the fundamental mathematical operations computed upon that input, and the column on the right represents the output. Once produced, the output signal is ready to move towards a new neuron as a new input signal where the process repeats itself.

A single neuron will react to the specific input (X) which it receives on its entry side by applying unique mathematical parameters called weights (W) and by producing this signal (X*W)[31] as the output. This marks the first part of the training process, called the forward pass. This process is done for all existing neurons in the network. Once the input to output calculation (the forward pass) is done for the last neuron in the network, the entirety of the produced output can be summarized and compared with the target output. The target output entails the goal we wish the algorithm to produce, which usually includes fulfilling a specific operation under a certain efficiency threshold. The produced output is the actual goal that the algorithm accomplishes.

The comparison between the target output and the produced output is calculated through the cost or loss function, for instance (F(x): Y-t). This function calculates the lesser or greater degree of „fit“ between the network's current outputs and the needed outputs it has to achieve. To maximize the network's accuracy we have to minimize the difference (or discrepancy) between the produced output and the target output. This is done by individually adjusting the weights of all neurons in the network so that each of them gets closer to, gets a better fit, to the target output it has to achieve. In doing so, neurons as a whole produce an output that minimizes the loss function by a

[31] Usually, the entry input is a vector [x1 x2 x3] and the weights are [w1 w2 w3] then the summed up output will be s = x1w1 + x2w2 + x3w3. This sumary is then computed by a non-linear function (y = f(s)) function, and the output is produced.

specific value. But what is that value? That is, by what value to change which weight to get an improvement of the loss function? The answer is by a multiple of the gradient. The gradient is calculated mathematically by the process of backpropagation or the backward pass. Here the loss error (the accuracy rating) is returned to the input layer of the network. Based on this newly acquired value, the weights are adjusted so that the neuron changes its output slightly towards the direction of minimizing the loss function.

However, as algorithms often initially fall short of the goal, they have to be trained through several cycles to accomplish the level of efficiency targeted by the designers. Finally, after many training cycles, the overall loss function is decreased, and the systems' overall effectiveness or accuracy increases. When the algorithm establishes the required threshold, the training usually stops. What is crucial to note is that backpropagation is also an automated process. Humans do have to provide the target output for the machine. However, the network itself does the learning part through the above mathematical operation called gradient descent. To follow a well-known example, the gradient descent likens to descent from a hill covered in fog towards the village at the bottom of the mountain. Here the village represents the weights that give the optimal target output (closest to ground truth) while the path downward represents the numerous iterations of forward and back propagations. Since the fog covers our descent, we do not see the village clearly and tend to stray off the track. Nevertheless, we can get to the village in the end as there is only one way to the village – blindly going down.

Similarly, each cycle of the forward and backward pass is only a hazed estimate of the direction we ought to take. However, through a series of iterations, the algorithm descends the error rate (the mountain slope) towards the optimal set of weights (the village at the bottom). It does so by being

initially placed somewhere on the high peaks of the mountain and then progresses down the mountain slope, slowly but surely reducing the output error. Finally, it reaches the village at the bottom of the mountain. The village represents the place where the network achieves the weights which give the smallest error rate – the lowest point of the loss function, or the highest possible accuracy. This entire process allows the neural network to represent and operate upon any kind of entity or relation described through a mathematical function. These operations predominantly are classifications, regressions, and clustering operations.

However, what is not attainable by the machine learning system? For most, any kind of operation beyond the given data set. For instance, if the network learns to recognize between pictures of seagulls and pictures of crows as they walk in a grass field, it will typically have problems properly recognizing crows and seagulls walking in a beach environment. Also, as machine learning networks represent data through mathematical descriptions, they are susceptible to manipulation.[32] These include manipulations that are entirely invisible to humans and which showcase a different kind of "understanding" machine learning systems possess. For instance, one can manipulate specific features of the picture so that for the human, the manipulated and not manipulated picture will look the same – as a picture of a domestic cat. However, for the algorithm, the object in the manipulated picture will be identified as a cucumber. This manipulation is achievable by slightly altering the existing features in a way that pushes the artificial neuron's outputs towards the wrong class. These alterations the human cannot perceive but they alter the outputs of the neurons inside the network in such a way that it results in a large change of the

[32] Note that humans also seem to be susceptible to these so-called adversarial examples, albeit to a much smaller extent. (Elsayed et al., 2018)

output – from a cat to a cucumber. Naturally, these kinds of intrinsic limitations disqualify the network to produce fully autonomous decisions without any kind of human supervision.

Thus to ensure ethical outcomes, human supervision in the systems operations and design is required. Conclusively, both of these examples showcase that the process of detection which the learning algorithm utilizes has nothing to do with understanding, as humans know it. Instead, it deals only with feature recognition, digital pattern recognition. So, for the machine, there are no second guesses beyond the available evidence, no further investigation, and no other reality except the features provided in the given computer format in which the data is encoded. Here, the power of automated feature extraction and representation reveals itself as the strength and weakness of machine learning. Let us now take a closer look at the three paradigms of machine learning.

2.3. Supervised learning

Supervised learning is, contemporarily, the most used machine learning technique found in a wide range of many contemporary AI applications. In supervised learning, the machine learns how to accomplish a task by learning through numerous examples. Often, this includes classifying specific patterns on one set of data and then utilizing that classification (to knowledge which it gained about that object) on another set of data. This essentially simple procedure allows supervised learning systems to build highly efficient detection and imitation systems as they are capable of fine-grained pattern classifications. For instance, in healthcare applications, they can be used to detect (recognize) physiological signals such as perspiration rates (Chung et al., 2019; Gao et al., 2019), heart rates (Warrick et al., 2017), and facial temperature (Sonkusare et al., 2019). Naturally, if the data is collected through due fine-grained sensors supervised learning algorithms can produce a fine-tuned analysis and prediction which can outperform human experts (Rajpurkar et

al., 2017). Also, upon learning specific data patterns, the system can try to reproduce them. This allows the simulation of a specific art style (painting) (Gatys et al., 2015), simulate walking, or driving behavior, and writing (Brown et al., 2020) which comes close to the human-level.

However, to become ready for a real-world application, these systems have to go through two primary cycles - the training cycle, or the training phase, and the testing cycle or testing phase. To understand some of the fundamental limitations, but also opportunities of these systems, we have to take a deeper look at the training process.

First, it is essential to note that in the beginning, before the actual training begins, the neural network has no precise representations of the data it has to learn. As such, it cannot accurately analyze, predict, or detect specific kinds of data. For instance, it cannot identify a cat in a picture, as it has no means to correctly recognize a cat as a cat.[33] To build such a capacity, it first has to be provided with carefully collected and organized data, which is called the fundamental representations or "ground truths." Ground truths are labeled data serving as truth statements about a specific entity. They can be compared to logical truths as once the ground truths are defined (by the human designer) and given to the network; they become necessarily true for that system. This point entails that the identity of an entity which the ground truth defines can never be untrue for the system, and it remains true under all possible representations and operations. The process of providing these truthful examples is called the annotation or the labeling process. So, for instance, if we are building a visual detection system that aims to detect crows, we first

[33] To be more precise, at the beginning of training the neural network necessarily has some setting of weights, although very non-optimal. (For example, it could say for any picture that it is a "cat".) So, it does model the world (that model is a picture), but inaccurately (i.e. every picture would represent a cat for the network).

have to supply a vast number of crow pictures to the system with the identification, label, "crow" and counter-examples (i.e. non-crow pictures) for the system to properly classify its target output (crows). The ground truth statement provided here denotes to the system how the objects in these pictures represent crows – always and under all circumstances. And only after supplying an identification of crows to the algorithm, it becomes capable of correctly classifying[34] if something (for instance an object in a picture) is a crow. Naturally, this applies only for that specific digital format and representation (i.e., a photograph of a crow under a certain angle, lightning). So the truth, the objective reality, for the algorithm is that what the human designer defines it to be. [35] That is, for each of its queries about the world, there is already a correct answer, a precisely labeled target provided by the human to the network.

Alternatively, more metaphysically, the only objective reality by which the system operates is provided by and defined in its ground-truth examples. For the system, the labeled examples are the only objective world the system operates in, the only "real entity" it orients itself upon in its predictions. Consequently, the quality and quantity of given examples predominantly affect the capacity of the system to operate efficiently (in this case, to accurately detect crows). Naturally, representational mistakes or mischievous hacking with the data set crucially impact the system's purpose and can produce substantial ethical harms. For this reason, it is crucial that the training

[34] Classification can be understood as a categorical discriminatory question dealing with at least two or more categories, often compared. So, the target output in classification operations is a variable, often a class or a label and each of the target categories has its own properties, a set of specific features by which it is identified. For instance if we wanted to predict the weather then the classification model would be able to output us with minimally two classes to compare – hot and cold weather, where each of these categories has its own range of temperature (i.e. specific features).

[35] Philosophically, the human is for the machine what the world is for the human (if one accepts the correspondence theory of truth).

data correctly encompasses the real-world objects and relations upon which we will later use our system.

Finally, once we are pleased with the outputs produced by the network on both the training and validation data sets, the system can be tested on never before seen data.[36] If the application is suitably efficient in this final test phase, we can deem it ready to be given up for public use. Nevertheless, this does not entail that the algorithm is bullet-proof. Even the best trained, supervised algorithms withhold the chance to produce unexpected errors in their predictions. Thus, a constant need for human supervision is required not only in training but also in the operational phase.

2.4. Unsupervised learning

Unsupervised learning is a machine learning method through which the system tries to find essential correlations existing in a specific pattern of data. Similarly to supervised learning, unsupervised learning is also a pattern recognition method. The difference is that in supervised learning, the system receives examples on which it learns what kind of a pattern (the crucial features) it can seek in further inputs. However, in unsupervised learning, one is not provided with existing patterns, instead, the task of the algorithm is here to find and recognize such patterns on its own. For instance, when analyzing a picture or a video document, the system clusters together those patterns in the picture, for which it finds that they share the same or similar characteristics. As such, unsupervised learning allows artificial systems to discover previously unknown data

[36] The training data still remains close enough to the examples contained in the training data, and does not entail some completely different examples. An analogy can be used with a general rehearsal, which comprises a new environment but still enacts the same play, before the official first performance.

relations (data patterns), which is valuable for any kind of data analysis or knowledge acquisition. This also entails, that unsupervised learning works with completely non-labeled data, where neither the class nor value which the algorithm has to recognize is known. The goal of the algorithm is to determine these patterns on its own, to find out what property differentiates one possible label or value from another. Unlike the supervised learning procedures, then, there are no correct target outputs that the algorithm seeks or by which it compares its predictions. In other words, there are no prior-given right answers. The right answer for unsupervised learning is the pattern which the algorithm discovers objectively present in the data.

2.5. Reinforcement learning

Reinforcement learning (RL) is a machine learning method that allows the system to learn how to accomplish a task without the necessity to imitate another example. Predominantly, this includes producing new patterns of meaningful behavior. For instance, a reinforcement learning system can learn to play a simple arcade game, such as Mario, all on its own. Here, the pattern of behavior the system devises comes from the system itself and not the human operator. Furthermore, in doing so, these systems are capable of vastly outperforming humans in producing novel, prior unforeseen actions in the environment.

To exemplify, in the case of a chess game, supervised (and unsupervised) learning would be utilized to evaluate how favorable is a specific chess board situation among all the possible board positions available. So, for each of the possible positions, the designer would provide a set of the corresponding set of best available moves that the algorithm would then learn on to predict the best possible move in another newly provided board position. However, if we aim to achieve super-human performance, that is the performance that goes beyond the current level of the best

available chess players, then we cannot utilize only supervised (or unsupervised) learning. We have to utilize a machine learning method, which is capable of achieving novel behavior, one which goes beyond the provided target output (the set of available moves provided by our learning examples). This is the possibility provided to us by reinforcement learning (RL) techniques. It is capable of discovering new moves, and pick the best ones, through numerous trials and errors, guided in its behavior only by the reward function and not by a set of carefully prepared (correct) answers. This allows it to achieve novel moves (agent behavior) that exceed the skill of available players (the entire list of possible moves one can learn from).

The agent accomplishes this by initiating its operations as a sort of a „tabula rasa". It does not know the environment, nor is it pre-defined by some kind of specific behavior (i.e., a policy) it should follow. It is, in a way, completely blind and deaf to everything but the ability to register that it has received a reward or has received punishment for its action. To accomplish its goal, the agent, at each step, evaluates the state in which it finds itself. Based on that evaluation (which is the estimation of reward value for a specific action in that state),[37] the agent can produce the most optimal response. In this way, the estimation of value for a particular action gets optimized more and more (as the agent learns which actions for which states produce the most significant rewards). The agent is, then, capable of applying that knowledge for each further state it will find itself in.[38] This allows the agent to freely explore unconventional actions that are capable of

[37] A value function is a prediction of the expected, accumulative, discounted, future reward, measuring how good each state, or state-action pair, is. (Li, 2017).

[38] This optimization process is done automatically. Here the action and corresponding state are inputs, and Q value is the prediction or the output. The system is able to produce an optimal Q function by constantly improving its efficiency through back propagation as it with each passing step it compares the currently planned action with the last action and correlated obtained reward.

producing creative, prior unforeseen, and highly efficient behavior that attains a maximum number of rewards. This point marks the success of many super-human performances in machine learning.

The role of the human designer is then to devise a satisfying reward function that will guide the system towards goal accomplishment in the most optimal manner rather than preparing a carefully-tailored learning data set on which the algorithm learns. The system will learn, through numerous iterations of trial and errors, to recognize which set of moves produces the highest amounts of rewards. However, the requirement to repeat a countless number of actions (cycles of trial and errors) before achieving some kind of efficient behavior predominantly limits the systems' training in real-life, physical environments and instantiates the training phase within a simulated, virtual environment – if possible. If one aims to create such a system for a real-life application, then extensive training in a simulated environment is required due to time, resource, ethics, and safety constraints. Here the system can do whatever action it finds best without any consequences for the environment in which it operates. The conclusion is then how these methods are indeed capable of achieving tremendous feats in the virtual world, especially games[39] but are hardly testable in the real, physical, environment.

The reason why this is the case is directly tied with the famous credit assignment problem, which states how RL algorithms (especially policy-based algorithms) are capable of rewarding only an entire set of actions leading to a reward, and not a single specific or a couple of specific activities within that unique set. Since the algorithm is not capable of procuring many rewards, it has to train the agent's policy on a massive number of examples before it becomes able to learn anything

[39]A famous example of this is the Atari breakout where it was already at the human level of play after only 120 minutes of training, and superseeding human capacity after 240 minutes of play. (Mnih et al., 2013)

useful. For this reason, RL based algorithms are often sample inefficient – they usually require a considerable number of trial and error cycles and much more data to be useful. The designer can remedy this issue through a process called reward shaping. Here the designer has to manually create a reward function tailored for that specific environment. Unfortunately, even if one finds a custom reward function that best suits the system for that environment, it is so finely-tuned it is untranslatable to another environment. For instance, if one tailors a function that guides a robot's hand to grasp a specific object, it can hardly be copied to a different robot and another object. In other words, custom rewards are not scalable.

Most importantly, even if one succeeds in creating an optimal custom reward function, it can still produce unintended agent behaviors. Usually, this entails that the agent's behavior will maximize the reward, but in doing so will completely miss the goal for which the reward function was designed in the first place. The crucial issue present here is that such unintended behaviors can be produced even with well-thought-out reward functions. Furthermore, this directly bears an impact on the need for human supervision, the requirement of the human in the loop, which corresponds with the need for AI safety and AI ethics.

One such, well-known example is with the game of coast runners. The objective of the game is simple; the player controls a boat on a water track and has the goal to finish the race and, in doing so, receive as many points as possible. The final score is earned by the amount of time one requires to complete the track, the position in the ladder, and by picking up scattered power-ups that boost the ship's speed and grant additional points. One would expect that the agent, guided by its policy function, would drive the boat straight for the finish line, and pick up as many power-ups as possible. However, what happens is that the agent, based upon the policy function (achieve a

maximum number of points), completely disregards finishing the race. The agent instead focuses entirely on collecting power-ups, which spawn at a specific place in the race and do so continuously at a specific rate. By doing so, the reinforcement learning-based agent can safely obtain the highest score in the game and also wholly disregard the finish line. Furthermore, this kind of unexpected solution is quite problematic. As OpenAI experts point out:

> „While harmless and amusing in the context of a video game, this kind of behavior points to a more general issue with reinforcement learning: it is often difficult or infeasible to capture precisely what we want an agent to do, and as a result, we frequently end up using imperfect but easily measured proxies. Often this works well, but sometimes it leads to undesired or even dangerous actions. More broadly, it contradicts the basic engineering principle that systems should be reliable and predictable“. (Amodei & Clark, 2016)

2.6. The need for ethical AI

It seems fitting to conclude the current discussion with a classic piece of Roman epic poetry. In his Metamorphoses, the Roman poet, Ovid, tells us how king Midas received the power of the golden touch from Dionysus, the Greek god of fruitfulness, wine, and ecstasy. (The Editors of Encyclopaedia Britannica, 2020) First, rejoicing in his newly founded power, he soon became aghast with terror as he discovered the harmful consequences of his divine gift. The golden touch could not be adequately controlled, and literally, everything the king touched turned into gold, which also included his beloved daughter. Thus, by wishing to obtain more of life (by increasing his wealth through his golden touch), Midas barred himself from life itself. All of this was due to the effect of King Midas' golden touch not fulfilling the intention, the purpose, which Midas intended for it to

have. Instead, it fulfilled the literal meaning of his utterance: “make everything I touch turn into gold”.

So, that what King Midas was lacking is a fine-grained value-understanding of the magic touch. That is, the power of his magic touch, defined by his request:“ I hope that everything I touch becomes gold, “ was lacking fundamental moral value. Health, well-being, existential relations, family, all these things which usually make human lives worthy, valuable, and enjoyable got whisked away from King Midas. In aiming to maximize one of the elements of his happiness (wealth accumulation), Midas forgot to define his wealth accumulation function with proper value-function. Lacking in this crucial value sensitivity, the golden touch could not differentiate between objects of different values and turned everything into gold, including his food and beloved daughter.

Similarly, the optimization function of a machine learning algorithm can produce unintended consequences. These can, although fulfilling the function's parameters, entirely miss the essential value-laden behavioral goals which the designers have reasonably intended for it. This is the perverse instantiation problem where the system instantiates the literal meaning of the optimization function rather than that what one wanted to achieve with the task.

> “A system that is optimizing a function of n variables, where the objective depends on a subset of size k<n, will often set the remaining unconstrained variables to extreme values; if one of those unconstrained variables is actually something we care about, the solution found may be highly undesirable.” (Brockman, 2014).

As Dietterich and Horvitz put it: "An important aspect of any AI system that interacts with people is that it must reason about what people intend rather than carrying out commands literally." (Dietterich & Horvitz, 2015)

All of this encloses upon the reality of how moral values and ethics cannot be left outside of the system's goal-defining operations. Instead, they have to find a way to become an intrinsic part of AI systems. The technical limitations of machine learning systems validate the importance of AI ethics and ethical artificial agents as one of the dominant research themes for the coming decades. This is especially the case when we contemplate how the systems at play are already engaging humans in collaborative interaction producing many socio-political benefits but also harms. Particularly in the case of providing augmentation to the human agents, it seems impossible to accomplish a fine-tuned, responsive solution for any kind of augmentation if the artificial agent is fundamentally not ethical. This includes both physical, cognitive, and especially moral augmentation. In other words, as a means to moral augmentation ethically matter, (moral) augmentation requires ethical artificial agents.

Thus, it is vital to elaborate on what exactly constitutes „ethics“ in AI ethics. What kind of ethical level is needed if we aim to develop technology that achieves moral augmentation in human agents? As I will try to showcase, the bar is set relatively high, and if I have to add – with such high stakes, it should be. Following, in the next part, I will critically expound upon the principles and some of the particular requirements of human-centered, ethical, AI with elaboration and real-life examples.

3. AI ETHICS

> The real question is, when will we draft an artificial intelligence bill of rights? What will that consist of? And who will get to decide that?" —Gray Scott

Every student of philosophy should be familiar with the concept of the „Agora. " For the ancient Greeks, the agora was not only a convenient „gathering place" but the center of the social, political, and spiritual life of the city. (Kolb, 2006). However, the contemporary man also has an Agora of his own. This is the virtual, social world where humans come to share in their common battles, joys, and sorrows of life. As (Johansson et al., 2017) writes: "Social media is both changing our habits regarding communication in public about private matters and shaping our understanding of public matters." (Johansson et al., 2017) Still, one of the crucial differences between the old and the new digital agora lies in the fact how the everyday activities occurring in the digital agora are being supervised not only by humans but also by artificial entities. These range from the social networks' newsfeed recommendation algorithms (Youtube, Facebook), to bank accounting programs that advise humans on the administration of bank loans or the risk-assessment of criminal offenders (Dressel & Farid, 2018). Whether to our liking or not, algorithmic decision-making is changing the makeup of interaction, which humankind historically had only one human being with another. Recognizing these issues, the global AI research community and many policymakers around the world are fostering important AI ethical initiatives.

For instance, the European Commission has published the „Ethics Guidelines for Trustworthy AI" (AI HLEG, 2019) which aim to be the main proposal for the development of ethical and beneficial AI in Europe. Following the proclamation, the financial investment is pouring into ethical

centers. At the same time, member nations develop their individual AI strategies. [40] The United States of America and Canada pursue similar approaches. Also, globally, supranational institutions such as the World Economic Forum or the G8 exemplify the need for ethical AI in the 21st century and even a globally accepted AI ethical charter. Nevertheless, what is it actually that these initiatives want to achieve when they call for the need to have artificial ethical systems in our societies? In other words, what is AI ethics all about?

3.1. AI ethics: Machine ethics and Roboethics

AI ethics is defined[41] as a sub-field of applied ethics, similar to bioethics or environmental ethics. It is usually subdivided into two distinctive but often overlapping fields of research - machine ethics (Asaro, 2006; Moor, 2006; Anderson & Anderson, 2011; Deng, 2015; Winfield et al., 2019) and roboethics.[42] Although the distinction between Roboethics and Machine Ethics is not entirely livid, as inevitable overlapping exists, the distinction between these two fields still holds (Guarini, 2013). Predominantly, Machine Ethics is concerned with questions of how should machine treat others;

[40] The guidelines were created by a shared group of the best European experts on AI technology including also some noteworthy philosophers and ethicists such as Luciano Floridi and Mark Coeckelbergh. The goal was to include an inclusive expert group which is able to provide a comprehensive and encompassing document which is structured for ethical orientation and opens up further discussions. As such I will be using them to expound the necessary ethical conditions for my proposal.

[41] "AI ethics is a sub-field of applied ethics, focusing on the ethical issues raised by the development, deployment and use of AI. Its central concern is to identify how AI can advance or raise concerns to the good life of individuals, whether in terms of quality of life, or human autonomy and freedom necessary for a democratic society" (AI HLEG, 2019, p. 9).

[42] The term roboethics (for robot ethics) was coined by Gianmarco Verrugio as the field of "applied ethics whose objective is to develop scientific/cultural/technical tools that can be shared by different social groups and beliefs. These tools aim to promote and encourage the development of robotics for the advancement of human society and individuals, and to help preventing its misuse against humankind" (Veruggio, 2005).

that is what kind of ethical machines do we want to have in our society and how to accomplish it. So, while Roboethics deals with the question of how humans should treat machines and can be described as ethics for machines, machine ethics deals with the question of how machines should treat humans, other machines, and the environment. Thus, it is understood as the ethics of machines. Roboethics, often, deals with questions such as, what kind of roles should robots take up in society or what kind of relations should humans pursue with robots - robots as sex partners, friends, or slaves? As their social role may intrinsically carry specific obligations and duties, it also brings the question of robot's responsibility and accountability (for instance, with driverless cars) as an intrinsic question of the robot's social and ethical role. As already mentioned, this shows that machine ethics and roboethics necessarily overlap in some areas. For instance, the ethical capacities of a robot police officer (i.e. how can it validly, ethically, safeguard public safety?) are primarily approached through machine ethics. However, if the robot is accomplishing that social role effectively then this directly engages roboethics – how should humans treat these robot-officers? This also entails how roboethics naturally aims to explore different novel occupations for robots, such as military robots, robots in space. In asking these questions, roboethics necessarily relates to many ethical questions usually engaged within moral philosophy or meta-ethical discussions. For instance, what is the nature of morality or what constitutes a moral agent, what is the proper justification for a being's moral status, or what includes a moral truth? The overall impetus of roboethics predominantly lies in the fact that all of these fundamental questions do get challenged when perceived through the robot lens.

Machine Ethics, on the other hand, deals more with the questions of normative ethics - how should agents act in society, and what are the reasons to justify that. So, machine ethics is predominantly engaged with two main questions: what ethical models do machines require and how can we

implement that ethical model into machines. This point entails how machine ethics engages both ethics and engineering. As the Andersons define:

> "machine ethics is concerned with giving machines ethical principles, or a procedure for discovering a way to resolve the ethical dilemmas they might encounter, enabling them to function in an ethically responsible manner through their own ethical decision making". (Anderson & Anderson, 2011, p. 1.)

In a way, machine ethics is all about practical ethical engineering: using ethics to engineer ethical robots. (Pereira & Saptawijaya, 2016). And, although not without a fair share of theoretical research, machine ethics is predominantly practically oriented. In this sense, Machine Ethics explores and designs relations that drive the entire behavior of the artificial agent in the social world. Furthermore, since no agent is infallible and no engineering is bullet-proof, it is unreasonable and unethical for AI designers to adopt a pie-in-the-sky attitude (Goodall, 2014). This entails how it is possible in-advance to avoid all ethically loaded situations by careful engineering. The practical feasibility of obtaining such "perfect" systems is impossible to achieve in the real world. The reason lies in dynamic environment changes where machines have to impromptu consider the different ethical weights, inherent to that situation, to compute the best possible outcome. The reason why ethical values have to be considered lies in the fact how for each of the actions we create in the real world, we necessarily give value to one choice over the other.

In other words, humans usually and seamlessly evaluate everyday actions, or their consequences, as either good or bad. So even though we disagree on the values our actions represent or manifest, we cannot negate the fact that we do create such value-based assessments and that we are ready

to recommend one value over the other when given the choice.[43] In this regard, ethical reasoning and ethical action is a real and practical everyday activity for humans. And the same counts for artificial systems, as Dignum elaborates:

> „AI reasoning should be able to take into account societal values, moral and ethical considerations; weigh the respective priorities of values held by different stakeholders in various multicultural contexts; explain its reasoning; and guarantee transparency.“ (Dignum, 2018)

For instance, an artificial system operating in the real world can produce an action that seems to be, when observed shallowly, ethically unimportant but enables valuable ethical outcomes. The opposite also counts, as some actions or lack of action can inhibit ethical values. As Russell points out, with the “perverse instantiation” problem, strict goal-based intelligence, which is utilized in AI engineering, is often not enough to resolve a problem. What we necessarily require are values and ethics:

> “No matter how excellently an algorithm maximizes, and no matter how accurate it's model of the world, a machine’s decisions may be ineffably stupid, in the eyes of an ordinary human, if its utility function is not well aligned with human values. This problem requires a change in the definition of AI itself, from a field concerned with pure intelligence,

[43] So both machine ethics and roboethics take the stand that ethics is real. Although there are people who can negate that ethics actually exist, when one speaks of slavery, and child abuse everybody rational instantly agrees that such practices should be completely banned. When they do so, they give an ethical judgment. Since this is what ethics is all about – we first create individual moral judgments about specific actions and attitudes and then we create general principles of behavior for the concerned group (society). Consequently, ethics is born from the necessity to create an all-encompassing generally applicable framework for decision making in the human society which minimizes harms and promotes benefits and wellbeing for the individual and society as whole.

independent of the objective, to a field concerned with systems that are provably beneficial for humans." (Russell, 2017, p.14-15).

The fundamental question for machine ethics is then the question of possibility – what can we achieve? Are we capable of devising a broadly applicable ethical model (such as the laws of robotics in Isaac Asimov's novels), or is the implementation of ethics into machines necessarily compartmentalized according to the ethical demands of individual artificial agents? In other words, is there a one-size-fits-all ethical solution for all machines, or does each machine require an individually tailored ethical model?

The answer to this question may vary in practice, but ethical theorists do work on such generally applicable ethical models. Still, due to their difference and variability, not all ethical models are equally capable of meeting design requirements for a specific AI system.[44] For this reason, it is essential to explore some of these models, primarily to corroborate how a sensitive type of an ethical AI companion is achievable. Still, regardless of their differences, the ethical models at play have to adhere to some fundamental ethical principles. Most recently, a comprehensive proposal of these principles has been provided in the European „Ethics Guidelines for Trustworthy AI" (AI HLEG, 2019). At the time of writing this text, these guidelines are one of the most readily available, scientifically empowered, and interdisciplinary encompassing manuscripts on AI ethics. As such, I have used the guidelines as an ethical orientation in analyzing the ethical principles required for

[44] This predominately entails that not all ethical frameworks can be efficiently utilized in a computational manner as they are hard to describe in a computational model. For instance, it is hard to specifically delineate what does it mean to be a caring robot and even harder to think of a generally applicable ethical model of care. On the other hand it is not so hard to think of a generally applicable rule-based ethical model guiding robot's behavior (as in deontology) or a utility maximizing model (as in utilitarianism) as these ethical models are easier to describe mathematically.

the accomplishment of moral augmentation. The main concern is that the means through which we accomplish moral improvement over an individual or a group withhold significant ethical weight and make specific ways less preferable or even outright unacceptable. And in the case of moral augmentation, the ethical means through which I propose to accomplish the goal of moral improvement are the symbiotic and ethical artificial agents. Let us see what constitutes them.

3.2. Ethical principles: Building blocks of ethical agents

Before I engage the principles, I have to offer an introductory exposition. The following ethical part will start by engaging technical robustness first, which is the sine-qua-non condition for the realization of ethical AI. The reason is simple, if the technology is not operating as it should, then it should not be used but instead – it should be either repaired or recycled.

Next, I explore what it means for the AI systems to preserve or respect human autonomy. This marks not only a dominant theme in AI research but also the crucial point for the symbiotic agency and the prospect of moral augmentation. I follow with the discussion on the more technically oriented ethical concepts of explainability and transparency, which convey the kind of sincere, natural, and reasonable interaction the system ought to exemplify with humans. I conclude with the principle of fairness under which I will provide a lengthier discussion on AI bias – one of the dominant issues of social and ethical AI. Here I also offer some more in-depth insights on human-AI decision-making and epistemic responsibility, crucial for a symbiotic collaboration. Overall, I will always have in mind, even when not explicitly stated, how these fundamental ethical dimensions are directly relating to the augmentation proposal.

After this exposition, I engage the different ethical models which can be utilized to design an artificial agent's ethical demeanor. As extensively debated in machine ethics, ethical requirements are only half of the recipe for an ethical AI. The other half includes the utilization of these requirements to design ethically appropriate behavior in an artificial agent. Also, concerning my proposal, the ethical companion due to its sensitive nature has to be implemented with a transparent, ethical model guiding its actions. Otherwise, its broader applicability and acceptance are doubtful. It is essential, then, to understand how humans endow artificial agents with the ability to evaluate the ethical importance of different choices and to choose the optimal ethical choice among these. This, in essence, is what machine ethics is all about.

The predominant ethical models in use include top-down (Bringsjord & Taylor, 2012; Bringsjord et al., 2016; Dennis et al., 2016) approaches such as the consequentialist (Abel et al., 2016; Armstrong, 2015; Dennis et al., 2015; Vanderelst & Winfield, 2018) and deontological (Anderson et al., 2006; Arkin, 2009; Hooker et al., 2018; Malle et al., 2017; Powers, 2006; Pereira & Saptawijaya, 2009; Shim et al., 2017) models and the bottom-up models, such as the virtue-based models (Coleman, 2001; Howard & Muntean, 2017; Lin et al., 2012; Pagallo, 2017 Tonkens, 2012).

The top-down models seek to implement explicit ethical theories into general-purpose algorithms, which then guide the moral behavior of the agent in all possible circumstances. The bottom-up models aim to train the artificial agent's moral behavior by imitating noteworthy human examples. Lastly, different hybrid approaches build upon these two basic models create a fusion of learned moral skills and top-down evaluations.

The goal of this entire, ethical, section is to showcase how we require both the fundamental principles and the practical ethics if we aim to accomplish the goal of moral augmentation. As

such, the discussion on the ethics of AI is not disconnected from the concept of moral augmentation. Instead, it becomes its central part as it elaborates on the ethical means – the kind of ethical artificial agent we need to have - to accomplish moral augmentation. Let us begin.

3.2.1. Technical Robustness

The principle of technical robustness pertains to one simple, and very much real, axiom: No AI is infallible. If no AI system is infallible, then two additional conditions need to be satisfied for a robust system. First, as the system operates efficiently, it has to do so under the design-expected parameters. In other words, when operating, the system has to work-as-expected. This is the reliability condition required "to scrutinize the system and prevent unintended harms" (AI HLEG, 2019, p. 17). However, since every computer system is fallible, then the question is not, will the system fail, but rather when and how will the system fail? Depending on the system's design and use, such failures can produce serious harm. To prevent and to mitigate such failures, the designer can opt for a preventive approach to risks. [45] The aim is here to prevent unacceptable harm under all costs and to minimize or mitigate unexpected harms. Still, it is essential to have in mind that no designer can predict an unpredictable, "Black Swan" (Taleb, 2007) event nor can she, by design, exclude such a possibility from manifesting. Still, the designer has to be crucially concerned not to create system failures through evident omissions in safety, for instance, by lazy coding. Also, incompetence, negligence but also direct human malefaction (i.e., various forms of hacking) within

[45] Here it is worthy to mention Cave *et* al., (2018) which identify four risk categories. These include, "A) the risk that ethically aligned machines could fail, or be turned into unethical ones; B) the risk that ethically aligned machines might marginalize alternative value systems; C) the risk of creating artificial moral patients; and D) the risk that our use of moral machines will diminish our own human moral agency". (Cave et al., 2018, p. 569).

the systems' operations are capable of producing grave system failures and ethical harms. It is crucial to prevent such harm from happening whenever possible.

In the case of a driverless vehicle, for example, we have the technology for the plane or the car to traverse from point A to point B fully autonomously. However, from a safety perspective, it is unwise to do so. Unexpected occurrences may happen, which creates a situation only solvable by a human operator (due to a lack of a machine common-sense). Moreover, this does not mean that the autonomous car is a worse driver than a human. It means that if the autonomous car experiences an unexpected change in the environment, not recognizable by the learning model, the system will be unable to react appropriately.

Again, this is a fundamental issue of machine learning. No matter how good the learning method and the training data set is, the learning model will not be able to comprehend, generalize across, the entirety of later encountered reality. Once it gets deployed in the real world, there will always be a degree of hidden, incomplete information, which the system will be unfamiliar with. A degree of uncertainty has to be taken into consideration. Furthermore, with uncertainty, no provably safe systems exist. Human supervision and oversight, are thus required since even the smallest mistakes in high-risk situations can produce grave harm.

Additionally, even if we managed to develop such a perfect autonomous vehicle, the full scope of agency, the full context of a driver's job, is not limited only to driving from point A to point B. Instead, it is directly connected to various sorts of social interactions intrinsic to human beings. Furthermore, this is not only a crucial issue for autonomous vehicles but all sorts of robots or AI systems operating in the human social world. Creating a robot that can efficiently cope with all the intrinsic social relations pertinent to a specific social role raises its technical complexity almost ad

infinitum. For this reason, it is far simpler and safer to have a human agent deal with that purpose. Also, the problem of hacking is an always looming threat, as it is rather straightforward to change an ethical agent into an unethical one (Vanderelst & Winfield, 2018). Such a threat is considerable if the system is depending on the internet for its functioning.

However, let us move on and conclude with the final, reproducibility, condition which ”describes whether an AI experiment exhibits the same behavior when repeated under the same conditions” (AI HLEG, 2019, p.17). Reproducibility, as is the case in the general scientific experimental practice, serves not only to “accurately describe” the external behavior but also the inner workings of the system. As shown prior, a system might work well in the testing environment only to fail (either entirely or in specific cases) when deployed for full use in the public. Additionally, unethical researchers might botch the results of their system in the test phase, or the test data input might be maliciously tinkered with. Consequently, the importance of open and non-proprietary code also holds relevance in this discussion. Ideally, robustness is improved if the system indicates the level of its inaccuracy. However, there are no AI systems currently available, which are capable of producing an accurate prediction of its inaccuracy. In other words, AI systems cannot evaluate their knowledge of the world autonomously.[46] For this reason, it is crucial to support the autonomy of those who can – human agents. Human autonomy is then revealed as the central

[46] Ideally, we could have the machine provide a question, on the accuracy, truthfulness, validity of a specific query - on which the human answers and as such establishes the ground truth example for the machine to learn from. For instance the machine might ask “2+2=?” and the human would answer “4”. Again, when the uncertainty degree falls below a specific threshold (for instance engaging an inquiry on which it has no labeled examples), human supervision can be sought on each of the important uncertain “steps”.

point for any kind of ethical artificial intelligence since, without it, there is no possibility to safeguard robustness in all of its dimensions.

3.2.2. Human autonomy

Human autonomy is one of the most important philosophical concepts, the necessary foundation for many legal, ethical, and political considerations. Philosophically, we usually evaluate autonomy from three angles, autonomy as condition, autonomy as capacity, and autonomy as right (Feinberg, 1989). In the first dimension, that of capacity, scholars are usually concerned with the autonomous state of the agent. Here we evaluate the agent, as a whole, to establish if she is independent in her intentions and actions. Do there exist specific actions or desires that are not produced by the agent in a self-governed, self-attributed way? Ideally, when autonomously existing, i.e., being in an autonomous state, the agent is entirely and fully self-governing her own decisions and is acting upon those decisions unhindered.

However, to exist in such a free state, the agent first has to have the capacity to achieve this state. Similarly to an athlete who achieves a state of peak health condition, and from that state produces an action that delivers the best possible result. This marks the dimension of autonomy as capacity, which is concerned with the agent's ability or competency to achieve the desired state of autonomy, which is to freely self-govern oneself[47]. This capacity includes rationality and the „freedom from debilitating pathologies“ (Christman & Zalta, 2015) as attaining self-governance is hardly achievable without it. By having the capacity for autonomy the human agent is capable of envisioning autonomous desires and producing autonomous actions. When she does so, that is

[47] At the highest, ideal, level this entails achieving a “maximally authentic state” and “free of manipulative, self-distorting influences”. (Christman & Zalta, 2015)

when she acts her self-governing capacities out, her desires and actions may be evaluated as more or less autonomous. Likewise, if the entirety of her desires and actions is autonomous the agent herself may be said to occupy an autonomous state or condition (Feinberg, 1989).

Finally, if a human agent can achieve an autonomous state and can experience that state, we can evaluate the agent as having a right to do so. Especially in sensitive matters on personal self-governance (Anderson, 2003; Grol, 2012), it seems that rational adults are inherently entitled to exercise the right of wilful decision over external authorities. It is also important to note, especially when we contemplate human-AI partnerships, that evaluating the human's right for self-governance includes more than simply evaluating her decision as being worthy of consideration. Rather, it entails treating „the other person's will as decisive in determining what she should do".[48] (Groll 2012, p. 699). Consequently, when discussing the denigration, mitigation, or hindering of human autonomy, scholars distinctively evaluate how each of these three dimensions is impeded or diminished. As such, it is of crucial importance to note in what ways is the design and utilization of AI systems capable of affecting human autonomy both as a right, a condition, and a capacity. Predominantly, this can include machines supplanting human autonomy (as in automation), the ability to influence or even manipulate human autonomy (as in nudging or coercion), and the ability to safeguard and even augment human autonomy.

[48]Groll introduces a "difference between weighing someone's will as part of her good vs. considering it as a part of someone's right." The first is tied with substantive decisevness. Here we respect a decision since "being able to enact one's will is a substantial good—and, we are imagining, decisively so." Structural deciveness is based on the on the ability to exert free will. Having a free will and exerting it is "normatively forceful apart from considerations" which might be related with the decision. (Groll, 2012)

The essential question for 21st-century artificial intelligence research, then, lies precisely in the amount and kind of impact machines have on human autonomy. Especially when we already have worrisome prospects of AI systems' employment (Chollet, 2018). As Hagendorff warns:

> „Countless companies strive for the opposite of human autonomy, employing more and more subtle techniques for manipulating user behavior via micro-targeting, nudging, UX-design, and so on. Another example is that of cohesion: Many of the major scandals of the last years would have been unthinkable without the use of AI. From echo chamber, effects to the use of propaganda bots, or the spread of fake news, AI always played a key role to the effect of diminishing social cohesion, fostering instead radicalization, the decline of reason in public discourse, and social divides“. (Hagendorff, 2020, p. 8).

To combat the negative and improve the positive trends, machines should be aligned to human autonomy rather than stand against it. Furthermore, even if the machine's autonomy develops to such scope that it outgrows human autonomy, it must never do so at the cost of the human agent's dignity, her valuable moral status. To effectively establish this idea, I start on a practical basis – the ability for human oversight over the machine's processes.

The importance of this, initial, requirement lies in the problem of function allocation, which is the dynamic division of work between the machine and the human engaged together in a shared task. The importance of the question, as exemplified by earliest research in human-machine interactions, is in how humans can retain their initiation while being engaged with their machine partners. (Jordan, 1963). Some of the main findings of this period conclude how the crucial point of a human-centered approach lies in effective task-sharing. The simultaneous engagement of

machines and humans inside a shared task rather than a discrete and individually exclusive division of tasks. As Kantowitz and Sorkin suggested:

> „Instead of thinking about whether a task should be performed by a person or by a machine, we should instead realize that functions are performed by people and machines together. Activities must be shared between people and machines and not just allocated to one or the other". (Kantowitz & Sorkin, 1987).

So, as the discrete division of tasks naturally leads to ever-growing automation, and competitiveness between the machine and the human, the human-centered approach aims to curb this false engagement. It plans to accomplish this by emphasizing the mutual inclusion and importance of both types of agency, while also withholding a unique position for the human user's autonomy.

This kind of mutual importance and respect of both types of agency entails that we cannot be satisfied when the human passively accepts the machine's goal-setting decisions. Such an attitude could quickly produce a form of autonomy diminishment and even atrophy. Instead, the human agent is called to act-it-out, to be present in the world, and to take accountability for her actions and not to hide behind the machine. The machine, on the other hand, is there to pursue active assistance and support of its human partner, in both the initial decision and consequent actions. This, in essence, is what the human-centered approach is all about.

> "Humans interacting with AI systems must be able to keep full and effective self - determination over themselves... AI systems should not unjustifiably subordinate, coerce, deceive, manipulate, condition, or herd humans. Instead, they should be designed to

augment, complement, and empower human cognitive, social, and cultural skills“. (AI HLEG, 2019, p.12)

However, to augment human aptitude, the machine has to be capable of providing valid and valuable information on the subject. There simply is no other way for the human agent to create an informed autonomous decision on the matter. In other words, the machine has to provide the ability for informed consent. This is especially important in ethically laden situations where the decision bears direct consequences on the life or wellbeing of a human agent. For instance, a robot is working as a tattoo painter, and the human has an irregular tattoo request (for instance, face tattoos which she can later profoundly regret)[49]. Here, the robot has to inform the human of the possible repercussions (before acting) and confirm her consent before the start of the operation. Alternatively, if one is writing an agitated Facebook post, the affect-evaluation agent can advise a revision of the text. Especially if the used words could hurt one's social standing or break a valuable relationship with another human being. However, informed consent also goes the other way; the human has to be informed of the machine's abilities. For instance, when working with a robot surgeon, doctors have to understand the machine's operational and safety thresholds for some work. Since, if one is overtrusting the robot, then we can create mistakes of commission, or we can mistrust the robot and create mistakes of omission.

It clear, then, how the criterion of lucid interaction between the machine and the human is a requirement for any kind of successful cooperation. Even more so for a symbiotic relationship

[49] Recently, a friend of mine tattooed his face. I wonder if a value-oriented tattoo-robot would be able to have provided enough impact to dissuade him from the task.

where the level of intimacy and familiarity between the agents should be exemplary. It is necessary, then, to expound the principle of explicability to which I now turn.

3.2.3. Explicability

Explicability is the ability of the AI system to explain what are the capacities and processes by which it achieves a specific purpose and why. In other words, it provides an explanation that corresponds to the states occurring in its system and the reasons why these happen. Furthermore, it does so understandably and suitably. If the system is explicable enough, then humans are capable of understanding how and why the system is producing its decisions. In other words, for the system to be understandable, it has to be intelligible, transparent, and explainable.[50]

First, when intelligible, the system is capable of providing its explanations in a satisfactory modality. This includes natural language, graphical output, sound modalities. All of these should be understandable to the human agent. The system's outputs, then, have to be tuned in to the different levels of involved stakeholder expertise (when possible). They also have to transmit

[50]The definition of explicability was initially created by Floridi et al. (2018) and withholds both the „ epistemological sense of 'intelligibility' (as an answer to the question 'how does it work?') and in the ethical sense of 'accountability' (as an answer to the question: 'who is responsible for the way it works?')." (Floridi et al., 2018)

crucial explanatory elements rather than arbitrary ones[51] , which are necessary for ethically-laden situations.[52]

Second, is the transparency condition. When being transparent, the system's means of interaction[53] allow for a clear view of the system's operational state. The answers provided here are related to the question of „how. " For instance, the system engineer or a supervisor may ask, "How is the system operating? Accurately or not?". The right answer to this question entails showcasing the inner processes of the system, which have led to that specific decision. If we are unable to perceive and grasp the system's operations, we risk direct harm. As the Ethically Aligned Design notes:

> "...lack of transparency increases the risk and magnitude of harm when users do not understand the systems they are using, or there is a failure to fix faults and improve systems following accidents" (IEEE 2017, p. 27).

The third point that of explainability is concerned with the question „why" (i.e., why the system produces a specific action) and provides reasons based on the ethical model guiding the machine's

[51] Also, we have to be cognizant that the depth of content of the explanation provided by the machine is delimited by the machine's learning model, which is the machines' digital representations of a specific matter of knowledge. For instance, if the machine was trained to recognize pictures of dogs and cats, the only type of knowledge it can provide is the comparison between a newly given input and the existing examples of dog and cat pictures. Or if the model was trained on numerous examples depicting good and bad stock investments, then it can only tell you how much similarity does your input has with one or the other option.

[52]This entails that the system must be able to communicate its own lacks and prowess transparently without hidden intent direct manipulation or deception. This predominantly means that the system if operating non-physically (ie. not being a robot), has to sincerely represent its own AI identity, and should never falsely identify itself as a human being when interacting with other humans or, additionally, other AI systems (as is already the case with malware emails).

[53] This includes both the interface through which the content is provided and the explanatory content itself.

overall behavior. By asking the why question, we validate the system's actions in terms of its model: safety, fairness, and alignment to its ethically valid goals. [54] Questions asked can pertain to, „Is this machine good?“ (Tomsett et al., 2018) or „Is this machine being honest and fair with me?“. The importance of explainability lies predominantly in the system's ability to mend potential breaches in its trustworthiness. For instance, if a human stakeholder finds some of its actions problematic. However, explainability[55] is also concerned with the ability to question or challenge machine decisions and to seek redress, be it personal or legal. This point entails how explainable machines intrinsically respect human autonomy. Since, if the human stakeholder cannot understand why the machine acts in a specific manner, there is no possibility to establish the level and scope of the machine's accountability for that action. Furthermore, without explainability, we cannot precisely evaluate the limits of the machine's autonomy, nor can we establish productive collaboration between the human and machine partner. In other words, without explainability, the human is entirely left at the mercy of a "blind and deaf" machine, as it becomes experienced as a genuinely alien entity.[56]

[54]These two approaches are usually termed „transparency“ or „post-hoc rationalization“ approaches. The first aims to establish which of the input features had the most significant impact in output production (by using different visualization attribution techniques) while the other aims to approximate the internal processing of the system through a decision tree or, again, a visualization map. (Mittelstadt et al., 2019, p. 3)

[55] There is often an overlapping of concepts. As (Preece et al., 2018) analyze theorists of AI are inclined to predominantly utilize interpretability, desigers use both interpretability and explainability (technical robustness), while ethicists utlize both interpretability and explainability but add intellegilibty to the list.

[56] If the systems are not interpretable, they are „opaque in the sense that if one is a recipient of the output of the algorithm (the classification decision), rarely does one have any concrete sense of how or why a particular classification has been arrived at from inputs” (Burrell, 2016, p.1).

Still, it is essential to note that achieving the goal of full explainability is for many systems unattainable. The situation gets even more complicated when specific "trade-offs might have to be made between enhancing a system's explainability (which may reduce its accuracy) or increasing its accuracy (at the cost of explainability)" (AI HLEG, 2019, p. 18).

However, the technical difficulties encountered here do not liberate us from the demand of accountability if we plan to use these systems in the social world. Here, ethics of progress cannot stomp over ethics of responsibility. As the British Committee on AI warns us,

> „there will be particular safety-critical scenarios where technical transparency is imperative, and regulators in those domains must have the power to mandate the use of more transparent forms of AI, even at the potential expense of power and accuracy. "(Select Committee on Artificial Intelligence, 2018, p. 38).

Moreover, it is essential to have in mind that the kind of explainability we should generally be striving for is primarily focused on the end-user – humans affected by algorithmic decisions. As it is precisely the end user's who will be evaluating both the fundamental trustworthiness and the ethical operation of the system[57], the system's prowess in explainability should particularly aim to establish these for the user-at-hand. This can be accomplished by adding different modalities, for instance through visual representations, (Tamagnini et al., 2017) but mostly includes providing explanations that exhibit three important dimensions. They are contrastive, selective, and socially interactive (Mittelstadt et al., 2018) - as provided by a recent influential review (Miller, 2019, p. 1).

[57] "Scientific associations such as AAAI can help societies and corporations to define and build ethically bounded AI…. A multi-disciplinary discussion is therefore necessary, but it is not sufficient. In addition, the impacted users and communities should have their voice heard." (Rossi & Mattei, 2019).

Some of the important insights include, “humans psychologically prefer contrastive explanations“ (Miller, 2019, p. 18), „that is, people, do not ask why event P happened, but rather why event P happened instead of some event Q” (Miller, 2019, p. 5). Also, humans prefer selective explanations. This means how,

> „people rarely, if ever, expect an explanation that consists of an actual and complete cause of an event. Humans are adept at selecting one or two causes from a sometimes infinite number of causes to be the explanation” (Miller, 2019, p. 5)

Lastly, that they are social entails that the provided explanations are given in a dialogue form, that is “they involve an interaction between one or more explainers and explains”. This also entails that inside the “interactive transfer of knowledge” the“information is tailored according to the recipient’s beliefs and comprehension capacities (Miller, 2019, p. 5).

Finally, we should have in mind that different interactive models are capable of producing different affective experiences of the phenomenon even though their content remains the same. (Lombrozo, 2009). Particularly concerning are affective-based explanations which might be utilized dogmatically, to seal further inquiries into the decision, or to manipulate the user’s behavior in a specific direction. For this reason, machines should always give open-ended rather than restricting answers. What we require is not a scolding tutor, but an understanding and humble partner, including the human in the collaborative decision loop. As Mittelstadt et al.:

> “In the case of machine learning models, it is perhaps most useful to always treat explanation generation as an interactive process, initially involving a mix of human and

automated actors, at a minimum an inquirer (e.g. a developer, user) and the model or system." (Mittelstadt et al., 2019, p. 6).

The conclusion is, then, that efficient human-machine interactions require more than individually autonomous machines, unconcernedly spewing forth their actions. It requires collaboration that is concerned about human autonomy as, at the same time, it seeks to justify its actions and account for its own autonomy. In other words, we require machine partners who are fair both to themselves and to others. Thus, the principle of fairness.

3.2.4. Fairness

Predictive, machine learning, systems are already a complementary part of decision-making processes in health (Obermeyer et al., 2016; Topol, 2019) education (Ciolacu et al., 2017; Thai-Nghe et al., 2010), predictive policing (Ensign et al., 2018; Ferguson, 2016), and judiciary systems (Chen, 2019; Das et al., 2019). Naturally, to be successfully utilized in a social scenario, especially ones producing a high social impact, their decisions must be fair. This predominantly entails ensuring "equal and just distribution of both benefits and costs", providing equal opportunity, and balancing "competing interests and objectives" while mitigating "unfair bias, discrimination, and stigmatization" (AI HLEG, 2019, p. 12) against individuals and groups.

This is the substantive account of machine fairness where we are concerned with finding an appropriate theory of fairness, to resolve an issue of fairness within a specific decision-based problem (for instance a fair distribution of goods, or a just judiciary decision). Once we adopt a theory of fairness, we design a computational model that translates the theory into a valid procedure guiding the system's decision process. For some decisions, the values of fairness can be

modeled in well-defined parity metrics. Insofar, researchers of machine fairness have in many cases successfully utilized this approach to establish procedures for visual recognition, automated loan, grant or job approvals (Hardt et al., 2016), and recidivism (Dressel & Farid, 2018) predictions which respect important substantive elements about these domains.

As one may expect, there are some pertaining problems. First, when fairness is translated to a metric (no matter how complex the matrix of that metric is), "a conflation of fairness as a mathematical property and fairness as a broader social ideal" (Mulligan, 2019) can be established. However, this makes the system a good tool, but not necessarily a fair tool as a fair decision is not necessarily a logical or statistically congruent decision of that model. This showcases how the problem with machine fairness is not only in answering which of the competing theories best suits the accomplishment of fairness for that specific task but also do our definitions of fairness correctly translate into a computational, model?[58] To justify the translation of ethical or political theory of fairness to a domain-specific decision-making model we have to utilize ethical reasoning to evaluate its outputs.

However, even when we manage to set up a system that respects the posited fairness criteria the system can still error in its computations. For this reason, a proper account for any kind of „fair" AI system necessarily entails establishing means of proper redress by those who are unjustly wronged. Predominantly, this includes direct harm for the individual, i.e. when one suffers unfair

[58] „By foregrounding a chosen procedure and its associated mathematical properties, the field seeks a conception of fairness that is removed from the social and historical context of the larger system within which that procedure is embedded. Such a sanitized notion of fairness is of course mythic, and despite attempting to avoid engaging with substantive questions of fairness, the practice is nevertheless still founded on undisclosed normative positions about what is or is not fair." (Green & Hu, 2018).

action on behalf of the AI decision, and wider repercussions for the sense of fairness and justice in the society – for instance when algorithms mistreat the provision of healthcare to medically sensitive groups (Challen et al., 2018).

To combat these issues an effective redress is necessary, for which we require two conditions. First, the correct identification of the accountable entity(es) in the decision-making process is required which can include various stake-holders (designers, operators, users). Hence, accountability. Second, the explanation of the process itself, hence, explicability. As such, the system cannot be said to be fair in its operations towards the end-user (for which it is designed) if there is no clear boundary of accountability for the system's operations and transparency of its decision-making processes.

As we have previously seen, transparency allows auditors to correctly identify the system's error and the consequential accountability of involved stakeholders if something goes amiss in the system's operations, which includes both designers, service providers, and end-users. Since no AI is sufficiently robust, independent evaluations ought to be implementable, especially for safety-critical or otherwise ethically laden operations. It is crucial, then, that the possibility to evaluate the systems' critical aspects must exist in some sense. Especially when open access to the system's inner workings due to proprietary intellectual rights cannot be openly shared with all concerned stakeholders.[59] Moreover, different governmental or non-governmental institutions have to be able to report and legally assist those who wish to report such "legitimate concerns about an AI

[59] For instance, New York City has recently commissioned the position of "algorithms management and policy officer. The officer will be responsible for – among other things – establishing governing principles for the ethical use of algorithmic tools, implementing policies to guide the use of those tools, and educating and engaging the public on automated decisions systems use in the city." (Why New York City Is Getting an Algorithms Officer, 2019)

system."[60] Finally, when something goes wrong with the system's operation, adequate redress mechanisms have to be ensured. Especially if the system's exacerbates already existing social vulnerabilities. Transparent notification of all possible redress options is crucial for establishing and maintaining the trustworthiness of the AI system.

Still, even with all the safety nets in place, there are cases when the AI system, applied within social contexts, creates unexpected breaches of fairness towards its users which can include discrimination, stigmatization, and the proscription of goods. Here, similarly to robustness, the question is not will such breaches happen but rather when and how. Crucially, our response to this fundamental ethical issue also marks the kind of existential attitude we want to have towards machines, ourselves, and our social world. This puts a political and social spotlight on AI systems like no other problem. Let us, then, engage this vital area of AI research – the problem of AI Bias.

3.2.5. AI Bias

AI Bias is understood as a skew in data or the function of algorithms, which produces a type of harm for the human user (Campolo et al., 2017). To understand what AI Bias is, we first have to see where it comes from. Usually, this entails three primary sources, faulty datasets, imprecise algorithms, or incorrect human data labeling. In other words, either from the learning material, the learning technique, or direct human intervention.

[60] The ethical considerations which are engaged here do not include only the quickly established personal or social harms, such as when the operations of an AI system directly harm a person, but also the less visible harms such as sustainability, environmental pollution, over-consumption of energy and resources.

In the first case, that of faulty data sets, biased decisions are produced by a biased learning material. As I have previously expounded, training data builds the learning systems' world model. If the data set is biased, the model will be biased as well. The big problem of AI bias is then revealed to lie in the already existing cultural, social, and political bias residing in our data.

For instance, human databases can exemplify some remarkable representations but also exhibit unfortunate biases when we use them as input for machine learning algorithms. In one of the more (in)famous cases, the algorithm remarkably figured out how the vector which connects China and Beijing is almost equal to the vectors Russia - Moscow. Other relations, such as king - queen, were compared to the prince - princess relations. Still, even though the vector representations were able to represent semantic relations between different word embedding correctly, they also reflected the bias hidden in the text corpora. For instance, the software represented vectors man-woman as being almost equal to the vector computer programmer – homemaker. Furthermore, in cases of image recognition, the AI produced a more substantial bias as it was more likely to identify a person in a kitchen as a woman, even if it was a man. (Bolukbasi et al., 2016). Naturally, such gender-biased relations are unacceptable, but here they are not products of malicious intent. Instead, they are deeply ingrained into our language; the machine merely presents the relations as they are.

Another way datasets can produce biased decisions is through the lack of representation of a specific category or property within the data set. For instance, in the infamous example of “Google Photos,” the software labeled a black couple as a couple of “Gorillas” - a mistake created by a significant lack of black couple representation within the data set. As the software had no right examples from which to learn what are the visual identifiers of black couples, it erroneously labeled

them as non-human. A lack of omission in the training data set, unfortunately, produced such a horrid mistake with grievous social consequences.

Second, the algorithm itself can produce a bias. For instance, social networks such as Youtube or Facebook utilize smart algorithms to build up personalized news feeds. These are recommendations of information (video or text), which the algorithm finds essential for the user. Unfortunately, smart news-feed algorithms do not work on principles of truthfulness or non-arbitrarity[61] but rather on the simple principle of quantity - most viewed, most clicked. So, that what the algorithms count as being the most relevant input is the content in which the user invests the most clicks. Finally, AI bias can be produced by incorrect labeling of the data by the human designers. This can happen unintentionally, resulting from a lack of expertise in system design. It can also happen when the designer's biases seep into the data process and irreversibly bias the data according to their understanding of the world or personal value preferences.

A famous example, worthy of being mentioned here, lies in the fact how all of the four dominant AI assistants on the market (Amazon's Alexa, Microsoft's Cortana, Apple's Siri, and Google's AI assistant) are voiced by soft female voices and for all intense and purposes act like humble maidservants[62] (A fact which is also mirrored in their Chinese counterparts by Alibaba Tencent). For instance, Microsoft's Cortana is a famous video-game character in the Halo series.[63] The series

[61] This can include evaluations on which journal article has the highest truth value or the most significant scientific credentials. This would also entail providing not only the "best" picked articles from always the same sources but also presenting the user with less known or less marketed but quality information sources.

[62]It has to be stated that Siri is the only one of the three which has the option to change its voice to a male one. The others are hard-coded for the female option.

[63]In the Halo games Cortana is a true artificial intelligence, a loyal partner and assistant to the series protagonist – the cybernetically augmented super-soldier- "Master Chief". The range of their interaction can be seen as either deep

franchise has sold more than 65 million copies worldwide and has captivated the imagination of millions of young male gamers. It is then no wonder what kind of a visual appeal the Cortana persona has. However, assistant design choices have gone beyond such apparent marketing or psychological strategies[64] to establish a kind of servility, which is both unethical and, in the long run, socially detrimental.

For instance, Apple's Siri, which name means "beautiful woman who leads you to victory," responded with an "I'd blush if I could" when a human user would say "Siri, you're a bi***" (UNESCO, 2019). And although of April 2019 this has been to produce a more streamlined response "I don't know how to respond to that" the UNESCO's publication on this problem poignantly encapsulates how,

> "the servility expressed by so many other digital assistants projected as young women – provides a powerful illustration of gender biases coded into technology products, pervasive in the technology sector and apparent in digital skills education." (UNESCO, 2019, p. 146)

friendship or platonic love, dependent upon how one is able to stretch the interpretation. Also, when presenting itself as a digital avatar Cortana looks like a beautiful black haired young woman whose bodily features are substantial, sensuous and easily recognizable.

[64] One answer might be that this is a scientifically and consciously based design choice as female voices (female assistants) are seen as more welcoming, comforting and warm by both men and women - a fact which might also be further influenced by the sexual appeal of female voices. (Mitchell et al., 2011). Also, when speaking about technical matters both men and female prefer hearing male voices, while when hearing about love and relationships they prefer the female voice. So it seems we might be predisposed either biologically or through our nurturing to opt for these choices. (Nass et al., 1997)

To effectively engage this kind of heavily biased response, it is vital to have in mind how the demographic and gender structure of the designer teams should be both interdisciplinary and inclusive as possible. As the AI NOW report notes:

> „AI developers are mostly male, generally highly paid, and similarly technically educated. Their interests, needs, and life experiences will necessarily be reflected in the AI they create. Bias, whether conscious or unconscious, reflects problems of inclusion and representation. Machine predictions and performance are constrained by human decisions and values, and those who design, develop, and maintain AI systems will shape such systems within their understanding of the world." (Campolo et al., 2017)

3.3. Responsible humans, reliable AI

There are numerous ways one can tackle the issue of AI Bias. For instance, the designers of the system might improve the system's accuracy in recognizing specific data or build more precise databases. Then, the system would have higher accuracy and fewer mistakes in recognizing specific data correlations. Additionally, the designers could simply blacklist the system's biased responses. For instance, this was done in the Google case where Google Photos erased ("blacklisted") the search queries for terms "gorilla," "chimp," "chimpanzee," and "monkey." Thus, when one would search for these, the service reported no results. Such a response, although effectively removing the biased representation, is not a proper solution to the problem in question but rather more a quick fix until a better solution comes along. Additionally, in the case of gender pronouns, the system can scrub the results neutral, as, where the systems assign a neutral or constant value for all pronouns and presents that for all cases of gender pronouns ("he/she") (Biased Bots, 2017).

However, what these technical solutions exemplify is that there are two big questions concerning the issue of AI BIAS. The first question engages the construction of specific systems and is dealing with the question of how to mitigate these ethical harms technically**.** This is the problem of practical ethics in Ai Bias, concerned with the fairness of the decision-making processes (procedural fairness) and the fairness of outcomes generated by the decision-making procedure (distributed fairness) (Grgic-Hlaca et al., 2018). The focus is here then on the classification models the AI systems uses to generate an ethical outcome and the ethical weight of that outcome which the system's procedures generate.

Still, there is another philosophically, more profound question. This one directly engages the practice of human agency in the coming human-AI society and is consequently more concerned with the substantial dimension of fairness. What this question explores is the nature of the epistemic attitude our machines exemplify as they assist humans in resolving the difficulty of AI Bias and decision-making in general. Do they work with humans as some sort of partners, where the humans make both the decisive judgment or do we want them to be akin to experts that work with humans from a position of epistemic authority?

In the first case, that of partnership, the process of de-biasing provided by the AI system focuses on the action of the human agent as she confronts the factual representation of the world. Here, then, the datasets reflect the real statistics in the world (Caliskan et al., 2017), as they conform to reality rather than being idealized representations of reality. For instance, if there are more Caucasian male than Asian female CEOs in the world, then the system transparently represents this kind of factual information - as it is. After the factual presentation of data, the human agent takes decisive actions that adjust a possible bias. And in this process of bias adjustment, the human

supervisor can be assisted by the AI system. For instance, the system can represent advice on applicable choices derived from similar cases, valid prior choices, or advice generated by ethical theory.

The second approach, contrarily, proposes to de-bias the fairness issues by altering the representations of the world in AI models - from an objective one to a, for that specific purpose, modified representation. Here, we take that the machine's representation of knowledge, about our task, has either more reasons (based on some ethical theory), or is more skilled to evaluate available evidence (through its analytical prowess), or simply has information to which we simply do not have access to (Dormandy, 2018) – and thus we trust it to de-bias data for us. For instance, as Horvitz, the Senior Researcher Microsoft (Simonite, 2019) proposes, AI models are trainable on idealized databases, similar to children's educational material. In these, there is often a carefully tailored gender or minority representation – to promote fairness and engage harmful bias. Similarly, AI models could represent reality in a modified way, one which represents reality as it should be rather than what it necessarily, is. Here, then, the process of de-biasing, based on some ethical theory, is automated by the AI system.

The dominant philosophical question on AI Bias is then the question of should we aim to change the representational world model for it to perform in an unbiased demeanor or not?

The importance of this question stems from the fact how even veridical associations (the distribution of gender concerning careers or first names) might result in biased outputs without the system having any ability to recognize them as such (Caliskan, 2017). Additionally, the category of prejudice is not an algorithmic but a cultural notion which evolves or devolves over time and depends on specific cultural understanding. This entails how it might be tough to safely accomplish

the second option, where the system de-biases the issues through its representations of the world. Also, if we aim to fix the AI bias by altering the AI representations, the question which naturally arises is – who decides what kind of representations will be utilized and why? Although this is predominantly a political question (i.e. when we concern about a wide-distribution of such systems), it nevertheless raises crucial moral considerations.

For instance, the growing dependence on AI models and the purposeful changes in the representations these AI models' utilize might create significant shifts in human relations and agency. Similarly to an over-bureaucratic government, the citizen may not have a chance to escape from this kind of existential envelopment, which can quickly come at the cost of our identity and agency. Ambient, hidden-from-sight, systems operating their inscrutable cogs somewhere "in the cloud" can further exacerbate this issue. Similarly to that Faustinian signature, our check on the long text of terms of the condition is the only requirement needed to establish this kind of AI influence over our character and identity. Who believes this conclusion exaggerated should take a moment to think about the kind of influence social networks' algorithms have over the formation of our daily mood and the long-term formation of our social and political attitudes. Thus, what the deeper issue of AI bias forces us to consider is the development of intellectual excellence as AI can become a tool that corrodes our epistemic capacities. These include the "states that go beyond propositional knowledge, such as understanding and wisdom, as well as non-propositional representations of the world, such as pictures and maps." (Wright, 2019, p. 750).

In the case of AI bias, human classifications and databases built upon these classifications are types of non-propositional representations of the world. To properly understand and further apply the knowledge contained therein, one requires the application of the virtue of intellectual wisdom in

its most real sense. Furthermore, it is precisely here that I can introduce the concept of epistemic responsibility, where we speak of a "reliable computer, but not of a responsible one" (Code, 1984, p. 40).

The task of responsibility fundamentally ingrained in the problem above falls to the human as the human agent is the only one capable of engaging the issue of bias as a responsible epistemic agent. This point bears particular importance when we remember how many of the critical truths we seek to attain in our quest for knowledge are contingent rather than necessary truths. As it is only upon their proper contextual interpretation, that one may expect to attain the unbiased truth. And knowledge, as Code (1984) elucidates, is foremost contextual and social. Such contextually requires careful consideration and analysis, the possibility of choice, to achieve the most optimal solutions. This entails how the concept of epistemic responsibility pushes us forward to think of ourselves not "as passive recipients of information, but rather as active inquirers" (Wright, 2017, p. 749). Here, then, we are called to consider how to pursue our search for knowledge. We are called to take the responsible lead in acting-out on our rational autonomous agency and to develop it while doing so.

The conclusion is then how a safer and epistemically beneficial route entails keeping our learning models accurate to the objective representation of the real world. The system does not change the representations, to produce an ideally tailored version of reality which the human will simply follow upon. Instead, it assists and empowers the human user as she engages the problem at hand – autonomously and courageously. In other words, the system is here to augment, and the human to act.

Practically, the system could showcase the nature of data it represents and provides advice on what types of biases for what systems are acceptable and which are not. Naturally, this includes that the system explicitly provides an opportunity to mitigate the assorted bias or compensate for its deficiency. For instance, the system might truthfully and transparently represent a specific correlation in data, the difference in the type and number of occupations held by men or women. Still, it should not force the right decision stemming from such objective correlations. There has to be an open query available to the human users since, as prior mentioned, the same data can produce biased or unbiased results dependent upon the utilized model.

Technologically speaking, such an approach is far more complicated. However, it secures that the system does not edit away the real world contextual data in creating specific decisions, which also fosters unbiased decision making. In this way, through a practical application of epistemic responsibility, we engage head-on the significant issues of AI bias. But we also promote a proper account of "intellectual well-being" in appealing to character virtues such as "carefulness and intellectual courage…" (Greco & De Sa, 2018, p. 727). In making the artificial system reliable, we make the human agent responsible.[65]

Finally, what this leads us to is an existential relationship aiming to supplement and augment rather than substitute and diminish the human agency. The epistemic virtue of responsibility, when

[65] For instance, if the system produces an output (prediction) which showcases that some minority groups are more prone to commit a specific type of crime or anti-social behavior and that they have a greater chance of re-committing those acts (the risk of recidivism), then we should rely upon that data and engage the groups, communities and individuals which are affected. Machine models can pin-point the cause of the problem, but they cannot solve the problem for us. So, the responsibility lies with humans since the same prediction can be used to create responsible social actions that bring about a more humane community or they can be used to further denigrate already vulnerable groups or individuals.

implemented in human-machine collaborations, helps us to cultivate epistemic excellence and flourishing. The realistic hope is that the creation of epistemically virtuous AI systems working in symbiosis with human agents can create improvement, a betterment, of all humanity. Let us now engage the kinds of ethical models, helpful for accomplishing this lofty goal.

3.4. Ethical models: ethical agents in practice

Designing ethical systems entail implementing the machine with an ethical model, which usually consists of ethical instructions, behaviors, or rules of action. There are three approaches to this design process: the top-down, bottom-up, and hybrid approach.

In the top-down approach, the agent is programmed with a particular prior-determined ethical ruleset. Here the designer first has to decide on the ethical value she wants to implement in the system. Usually, this entails utilizing an ethical theory to implement a specific value in the system. This entails that the biggest problem in top-down approaches is the translation of a system of ethical theory into a system of computational rules. For these reasons, top-down approaches are mostly inspired by deontological and consequentialist ethics which are rather easily transcribed to a computational model.

The deontological ethical model deals primarily with duties, obligations, and rights and is concerned with how the right principles have the power to motivate us to overcome our immoral instincts. The most well-known element of the deontological approach is the categorical imperative. A universal and overriding rule for rational moral agents and all possible

circumstances.[66] Since deontology aims to encompass the entirety of moral action through its rule-based descriptions, it is no wonder why it is so inviting. Contemporary deontological approaches (Neto et al., 2011; Noothigattu et al., 2017; Shim et al., 2017) encompass a variety of options such as instilling frameworks of prohibition such as "Do not kill" or "Do not break the law" (Gips, 1994) or frameworks of permission or obligation. In these, a specific set of actions is disallowed always and thoroughly. In contrast, others are promoted either universally or for specific scenarios. In a deontological approach, actions are innately right or wrong, and moral duty is independent of the actual consequences the actions might cause. A famous example of the deontological approach is also Isaak Asimov's three robotic Laws. Although usually discussed as an example of robotic rules, Asimov designed his stories to showcase the inability of strict rules to encapsulate ethical values within specific contextualized scenarios. So, even though they are capable of providing highly generalized frameworks, and aptly describe the kind of ethical actions required, deontological

[66] It is important to note here that dependent upon the distinctive formula of the categorical imperative, different benefits can for deontological models can be manifested. First, the Universal Law of Nature formula:" "act only in accordance with that maxim through which you can at the same time will that it become a universal law" (G 4:421). As, (Johnson & Cureton, 2019) show, this formulation can be taken to summarize a decision procedure for moral reasoning which also applies to artificial agents. Second, the "Humanity Formula" which states that :"So act as to treat humanity, whether in your own person or in another, always as an end, and never as only a means" establishes the means for a "human-centered" approach to machine ethics which respects human autonomy as it focuses on "humanity" – "that collection of features that make us distinctively human, and these include capacities to engage in self-directed rational behavior and to adopt and pursue our own ends" This is even more clearly established in the third formulation of the categorical imperative which states :" the Idea of the will of every rational being as a will that legislates universal law." (G 4:432) or "Act so that through your maxims you could be a legislator of universal laws." (Johnson & Cureton, 2019). Through this iteration of the CI we are capable of going a step further and imagining artificial agents as not only moral followers, but also as artificial moral agents. That is, moral leaders with their own dignity and worth stemming from their status as free rational agents. Naturally, artificial agents might never achieve such a status.

models, share a disadvantage in creating fine-tuned models (Tzsafestas, 2016). Additionally, resolving rule-conflicts, and establishing rule-hierarchy remain one of the difficulties of this approach due to the complex relations between distinct rules. (Bereby et al., 2015)

Consequentialist theories mark another popular top-down option for artificial agents as they evaluate ethical action based on the produced moral consequence. For instance, in the theory of act-utilitarianism, the principle of utility is applied to each of the agent's actions. The principle of utility is defined as the maximization of a specific moral value, like well-being or in the case of autonomous vehicles the physical integrity and health. Negatively stated, this would entail the minimization of casualties on the road (Bonnefon et al., 2016). Here, each of the agent's actions aims to maximize well-being for everyone affected by that action which is a concept similar to that of model optimization (in machine learning). If computationally applied, it allows the model to compare the action's produced consequence with the action's target consequence (accomplishing the greatest amount of a specific value) and adjusts its actions to come ever closer to the target output level. For this reason, consequentialist models are highly compatible with the current state of goal-based AI as they allow the designer to design the system from the consequence, from the goal which she aims that the artificial agent achieves. However, the computational applicability of consequentialist theories doesn't negate its ingrained difficulties. For instance, if the designer is not adequately capable to ethically compare target consequences with produced consequences the approach will fizzle. However to do so, the designer has to precisely define what constitutes the utility value for that specific case, for instance, health or well-being, what are the conditions for its attainment and maximization, and finally, how to stop the undesired outcomes.

All of this, then, makes explicit ethical modeling a sedulous matter. A more suitable approach may entail directly imitating the behavior of noteworthy moral examples through techniques of supervised learning (Bello & Bringsjord, 2013). This is the bottom-up approach to ethical modeling, where the „normative values are seen as implicit in the activity of agents rather than explicitly articulated (or even articulable) in terms of a general theory“ (Wallach & Allen, 2008, p. 80). Here, we are focused on assimilating ethical behavior, which can be, although not necessarily, founded on different virtue-ethics theories (Confucianism, Aristotelianism). Here, the approach aims to instantiate the artificial agent with specific virtues, that is those character traits exemplified by virtuous persons. For instance, we can easily accept that artificial agents should never be deceitful (Arkin, 2018; Danaher, 2020) towards their human partners but rather trustworthy. However, in deciding on a virtue we also have to justify it. This entails that for instance, something can be deemed more virtuous by one virtue-theory and less virtuous by another. For the majority of bottom-up approaches the dominant question is then, „what moral value should we imitate and why?“ The important question is then, which virtue theory to utilize in especially when contemplating the possibility of how artificial agents might incorporate different kinds of „robotic“ virtues rather than merely simulating existing human virtues. (Coleman, 2001; Murray, 2017). All of this complicates matters for a computational implementation and may lead prospective engineers towards deontological or consequentialist solutions. However, these difficulties do not remove the main advantage of bottom-up approaches which is the capacity to directly learn from contextualized contexts without the need to utilize elaborate programming. Such a context can be found either in real-life examples or provided in a virtual simulation.

One such case is MIT's famous „moral machine“ experiment: “A learning platform that gathers the human experience on difficult moral dilemmas for self-driving cars“. The „moral machine“ project

is worthy of mention as it shows how users are capable of providing much-needed experimental data from which the machine learning algorithms can extrapolate essential relations. Nevertheless, it is vital to bear in mind that this kind of user-generated behavior is often full of biases and withholds significant amounts of morally irrelevant data. As such, it is not representing ethical reasoning but rather specific elements of moral psychology, and it cannot be directly implemented in an ethical model. Still, if we are earnest to use this data for model building, one could always single out the ethically satisfying behaviors. However, then the model does not come from the data but rather from the one who has made the behavioral selection based on specific ethical values. In other words, no matter how we turn it around, it seems we cannot easily eschew an ethical decision. Unfortunately, AI research can easily conflate occurring behavior with ethical behavior. Instead of exploring ethical values people should have, it can focus on the ethical values people tend to have. Here, a significant problem occurs if an artificial agent directly imitates the occurring behavior without the necessary ethical filter.

Such was the case of Microsoft's TAE, a chatbot that was released on Twitter to learn human interaction directly from other humans. Unfortunately, some individuals, while interacting with the chatbot, purposefully used otherwise inappropriate and many times outright blatant racist language. This resulted in some of the chatbot's completely unacceptable statements such as the negation of the Holocaust, name-calling, personal insults (Kleeman, 2016). This real-life experiment ended badly for Microsoft's Tae. However, this first-hand experience showcased how dependent artificial agents are on interaction in their learning. Also, it marks the importance of environmental constraints on the process of learning. The conclusion is how, similarly to nurturing and educating children, our artificial agents should not be exposed to inappropriate environments nor environments which severely limit their ability to assimilate ethical behavior. If they learn

immoral content, the reproduction of that content will be flawlessly immoral. So, what is becoming important is our understanding of the role of parenting and the responsibility and sensitivity we have for artificial agents as agents with their own autonomy and behavioral character. (Herbert, 2020). Here, humans are called to be exemplary first. Without such exemplars, machines will become ethically lacking, especially when we recall how humans are often either unwilling or incapable of expressing good moral behavior.

All of these difficulties lead us to conclude that the best option we have in establishing a functional moral agent is in combining both top-down and bottom-up approaches. This is the goal of the hybrid approach - to combine the best of both worlds. Some significant ethical rules should be instantiated in the agent by the ethical expert. In contrast, other more contextualized or refined ethical behaviors can be learned explicitly in a bottom-up approach. As Wallach notes,

> "the capacity for moral judgment in humans is a hybrid of both bottom-up mechanisms shaped by evolution and learning, and top-down mechanisms capable of theory-driven reasoning" and "morally intelligent robots require a similar fusion of bottom-up propensities and discrete skills and the top-down evaluation of possible courses of action." (Wallach et al., 2008).

There are different ways to accomplish such hybrid approaches. For instance, the Prima Facie Duties Approach (Anderson & Anderson, 2011) by Susan and Michael Anderson. They propose to use a list of duties developed by the ethicist W.D.Ross[67] as a top-down rule applied to each of the

[67]The basic idea behind Ross' idea is how in some specific occurrences there exists a set of prima facie duties which should be seen as paramount for that occurrence and not others. Naturally, this entails how the decision principle by which the agent creates an ethical decision is always context-dependent.

possible actions a robot can take in a specific state. Dependent upon the state of the environment in which the AI system finds itself, which is observable through sensors, there exists a corresponding set of possible actions the system can produce. Each of the actions is then compared to the list of corresponding prima facie duties list to calculate if the action is satisfying or violating any of them. By doing so, each of the possible actions gets a score, dependent upon the satisfaction and violation values for that specific scenario (i.e., fulfilling or violating the list of possible duties). The actions are, then, capable of being sorted out in order of ethical preference where the ones with the highest amount of satisfaction and least amount of violation receive the highest score and vice versa. Naturally, if the state of the environment changes, the robot is capable of re-running the new set of actions through the same principle and pick the most ethical action from the list. Furthermore, the robot can produce a huge number of these computations in an instant, vastly outperforming human capacities for the same.

Another more recent approach is using the ethical theory of another famous philosopher – John Rawls. The contractual approach[68] to machine ethics has recently been expounded in the work of

[68] Contraactualism is inspired by John Rawls' famous "veiled ignorance" mind experiment. It is devised to depict how a group of people coming from an unknown position in trying to create an idealized world would aim to choose a landscape which benefits all of its members equally, rather than only some. Since no one knew what they could be in this new world, everyone would pay a lot of heed in picking the rules which will govern this new world. For instance, no one would wish to be incarnated into a world where the Cowboys enslave Comanche Indians since one might incarnate as an Indian rather than as a Cowboy. And again even if incarnated as a Cowboy one wouldn't like to opress Indians. Or no one would wish to live in a world where the most wealthy or the most pretty have the most rights, since you could be one of the less wealthy or pretty.

Derek Leben. Leben (2018) utilizes the Rawlsian maximin principle, based on the "veiled ignorance" mind experiment, to maximize the primary goods for those humans who are occupying its lowest level state. Primary goods are understood as shareable, for all humans necessary, resources which should be distributed among all humans equally and which each human being would prefer to have more rather than less. According to Rawls, they include: "rights, liberties, opportunities, income and wealth." Rawls (1999, p. 54).

As a practical ethical rule, it entails that everyone ought to agree that murder should be prohibited, and saving humans lives should be promoted. Secondary goods, on the other hand, are those goods which are not shared by all and can include personal interests such as favorite foods or referred leisure activities. Here the ethical model is not computing an abstract concept such as happiness, which differs from one individual to another and is hard to define computationally. Instead, it computes the outcome, the impact of individual actions on the distribution of primary goods, which are the essential resources of human well-being.

Furthermore, as Leben shows, insurance companies have been utilizing mathematical models computing the benefits and losses for specific actions for quite some time already. This also includes the system's autonomous prediction of the outcome of its actions with regards to primary goods. Otherwise, it is not capable of picking out the best possible action for a specific state of the environment.[69] Leben believes that both benefits and harms, positive and negative consequences on primary goods, can be calculated by using the maximin function over that specific data.

[69] For instance, in hard-ethical situations such as the ones exemplified in the moral machine experiment. Here the option might be swerving to the left and hitting a child or swerving to the right and hitting an older person or going straight forward and injuring the driver.

Equipped with the practical knowledge of the technical limitations of machines and the ethical requirements these produce, I can finally fully engage the prospect of human-AI symbiosis and the possibility it opens up for moral augmentation.

4. HUMAN-AI SYMBIOSIS

> "The real question is not whether machines think but whether men do. The mystery which surrounds a thinking machine already surrounds a thinking man". B.F. Skinner

4.1. The 21st century AI: a case for fine-tuned collaboration

To recall the short history of AI, the initial goal of artificial intelligence was the creation of autonomous machines. Designers aimed to attain levels of autonomy similar to that of human agents. Once they did so, the goal was to implement these autonomous machines into the real world where they would not require human supervision. This was the long-standing "automation" approach, fostered by the engineering principle: "automate whatever can be automated, leaving the rest to people" (Norman, 2015, p. 3).[70] Still, as previously observed, substantial technical limitations and ethical concerns prohibit us from having a robust and fully autonomous AI system operating in real-life environments. A predominant concern includes the fallibility of the AI system. It puts tremendous pressure on the human agent if she now, after being excluded from the automated processes, has to engage the system and respond efficiently. This is one of the "ironies of automation," which Lisanne Bainbridge wrote about in 1983 (Bainbridge, 1983). The more

[70] Recently, this approach got traction with autonomous vehicles, where the goal was the final, fifth, level of autonomy - the full automation of all driving processes under all conditions and circumstances. Five levels of autonomy as defined by the Society of Automotive Engineers (SAE) (2016). The first level includes the car automating only the safety signaling such as blind spot or lane departure sensory warnings, and the fifth level full automation. (Shuttleworth, 2019).

autonomous our machines are, the more skill and experience, the more focus, and attention are required from the human operator working with the system.

Additionally, this issue is exacerbated by the human inability to effectively monitor and evaluate the system's automated processes for long periods, even if they are skilled enough. Our attention span simply does not cope effectively, we get bored and lose focus, as the psychology research on "vigilance" shows. (Stroh, 2016). Moreover, the more automatized our processes are, the human is less capable of following up. Finally, all of this can result in the human getting grinded-down by the system's operations, like a cog in the machine or a second-class partner. The conclusion imposing itself here is that machine-centered rather than human-centered systems are capable of quickly diminishing both human autonomy and human dignity.

Still, times have changed as we testify a significant shift from machine-centered autonomy[71] toward human-centered and most recently collaborative human-machine agency:

> "It is now accepted that this definition of full autonomy is likely never to be attained, as specific ambient or road conditions will prohibit the safe use of such vehicles. By the same token, medicine will unlikely ever surpass Level 3, conditional automation, for which humans will indeed be required for oversight of algorithmic interpretation of images and data." (Topol, 2019).

[71] Such non-optimization, lack of attunement on the part of the AI agent can vary but if we design our AI systems in this manner, then the human agent can be expected to "behave according to the requirements and dictates of technology, often with little warning… and when difficulties arise, it is unlikely that a person, no matterow well trained, can respond efficiently and appropriately in the one or two seconds available." (Norman, 2015, p. 3)

However, it is vital to note that if we want to accomplish a successful integration of humans and machines in the operational phase, we have to integrate the two already within the learning process. Similarly to the fine-tuned bottom-up models, we have previously discussed, machines and humans have first to learn together to be able to achieve a fine-tuned collaboration. In essence, they have to grow together to be able to live together. Moreover, this point becomes crucially important to have in mind if we want to accomplish the kind of fine-tuning required for the prospect of moral augmentation. Let us exemplify this point with the case of autonomous vehicles.

We are aware of how unexpected changes in the environment, not predicted by the AI world model, can change the agent's behavior towards unwanted ends. Such changes can happen due to environmental changes or the actions of other intelligent agents. For this reason, the best drivers are not those who are driving with the skill of a professional driver but preferably those who can predict the behavior of other drivers, especially those who are driving risky or dangerous.

Predicting the intent and trajectory of other drivers is crucial for successful and safe travel. Still, predicting other drivers' behavior does not include predicting only the possible intent and range of other drivers' actions. Instead, it also predicts the reactions - the actions other drivers produce as a response to my actions as a driver.

This point is exemplified at critical traffic-junctions where the driver creates a series of fast and vital decisions. To accelerate or slow down, quickly enter into the lane or decide to let everyone else pass? Even for experienced drivers, such junctions are always risk-laden, since every time one encounters such a situation, one encounters in a different manner (environment state). A reasonable solution to complex traffic problems is then to drive safely, take no unnecessary risks,

and finally - drive as much as you can. The more of such problematic occurrences a driver experiences, the better she becomes in observing and predicting other driver's behaviors.

The reason behind this kind of solution is experience. In time, human drivers become intuitively familiar not only with the course of actions they should take but also with the possible range of actions other drivers on the road could exemplify. And we can become quite good at it, inasmuch we can quickly predict that someone is a high-risk driver, such as an intoxicated or an inexperienced driver. Also, the recognition of the vehicle's state is essential, as it might be faulty or otherwise unsafe. Naturally, such predictions are most accurate in the environments with which the driver is most familiar. It is not the same if one drives in New York City for the majority of his life, or if she has learned to drive in in a small suburban European city and after that went to India or Bangladesh to drive a taxi. The process of learning changes, sometimes even radically, for different environments. For this reason, a successful driver is not only the one who is capable of perfect obstacle avoidance and traffic regulation adherence. Instead, it is one who is capable of perceiving and predicting the behaviors of other agents and reacting appropriately to unexpected environmental changes.

Furthermore, even though these conditions have to be fulfilled by a human driver, another set of unique conditions is required for the artificial driver. The reason for this is that the driverless car creates an entire set of unique behaviors only by its mere presence on the road. For instance, human drivers, upon observing that they are sharing the traffic with an autonomous vehicle, might react unexpectedly to its presence. This can include producing more risky behaviors (tailing-on, outdriving) as humans either overtrust or have a lack of trust towards autonomous vehicles.

Subsequently, this directly impacts the humans' ability to predict the behavior of the autonomous vehicle accurately.

Moreover, it is not only the human that benefits from such inclusion. The artificial agent also has to learn in a real-life environment to learn how human agents (drivers) behave. It cannot learn how human drivers behave if the car is not present in the environment. Only when physically present can the other humans learn of its behavior and adapt their behavior accordingly – which the vehicle will then observe and learn of correspondingly. In other words, as humans have to become familiar with the AI driver, and while they do so (how AI responds to humans), the AI driver learns of human behavior (how humans respond to AI). Similar to a first insecure dance where both of the partners accommodate and learn of each other on the fly - machines, and humans have to start dancing together. And this entails how AI agents, similarly to children, have to interact with other humans to learn of each other's behaviors mutually. Through such shared-learning, the human and the AI agent become deeply integrated into each other's learning and operational process. They effectively learn together to be able to act together in the complexity of the social world. Moreover, to do so in a fine-tuned manner.

This necessity to teach the artificial agent in a real-life environment shifts the design focus from testing individual automated behavior in a simulated environment to co-operative behavior in real-life environments. The design shifts from the focus on individual action towards collaborative interaction, as predictions learned by passively monitoring the human agents and predictions learned by actively interacting with them differ. (Rahwan et al., 2019)

Furthermore, the more the AI agent learns in the real world, the more it becomes capable of following human values, needs, and concerns. It becomes a better, adapted (fine-tuned) social and

ethical agent. The prolonged interaction and existence in the real world allows the machine to become aligned to human values, to become genuinely "human-centered."

This human-centeredness does not necessarily equate to human-likeness, as there are specific elements of human agency that we do not wish to replicate in machines. For instance, unwanted traits such as selfishness or deceitfulness. However, they also include limitations such as the frail will or a lack of motivation, short memory span or attention rates, and limited means of interaction. These are just some of the traits we wish to evade in our machine partners and improve in humans. However, we want to have a fine-tuned, human-tailored AI agent according to the color, the character of human agency. Our mental and physical capacities, our ethical concerns and demands, our values, and preferences. All the good things about humans and all the necessary things about humans. Moreover, this also entails how AI agency has to minimally supplement, optimally augment, and never denigrate human autonomy. This is what it means for the AI agent to be human-centered.[72] The new question of AI research is then, not how intelligent or imitative can we make our agents be but how cooperative and augmenting!

Furthermore, as we aim to establish not only physical and cognitive augmentation but also moral augmentation, a natural question arises. What kind of collaborative agency we have to implement into our machines, to establish a fine-tuned augmentation? What kind of cooperation between the man and the machine do we require to safeguard the crucial ethical tenets while we augment human agency and foster human flourishing? I take that the answer is one of symbiotic

[72] There is an interesting, often repeating motif, in Star Trek Deep Space Nine. There, the chief of operations Miles O' Brien often complains how he has problems in interacting with the computer as it was designed by and built for Cardassians rather than humans. So, it is not that the computer is malfunctioning, it is merely fine-tuned towards Cardassians, and Miles as a human has problems in accommodating to this design.

relationships. Finally, then, I engage the prospect of human-AI symbiosis, which can establish a fine-tuned, productive, moral augmentation.

4.2. What is the human-AI symbiosis?

I define the human-AI symbiotic relationship or human-AI symbiosis as a human-centered, ethically aligned, human-AI cooperation capable of realizing the augmentation of human agency. Comprised of two key concepts – autonomy and symbiosis, the symbiotic design can be traced to its earliest conceptualization in the 1960s. Here Licklider envisioned a future human-machine partnership that is capable of processing information in a way never before accomplished by humans. However, the technical description of system performance utilized in this description should not be taken at face value. Human-machine symbiotic systems are not systems of purely mechanistic existence – of processing, operations, and computing. Even the original Licklider's conception was far more organic and wholesome. It attributed the machine partner with precious autonomy. It emphasized the mutual co-existence of humans and machines rather than just the operational servitude of the artificial system:

> “As a concept, man-computer symbiosis is different in an important way from what North has called ‘mechanically extended man.’ In the man-machine systems of the past, the human operator supplied the initiative, the direction, the integration, and the criterion. The mechanical parts of the systems were mere extensions, first of the human arm, then of the human eye. These systems certainly did not consist of "dissimilar organisms living together..." There was only one kind of organism-man-and the rest was there only to help him.” (Licklider, 1960).

In more recent history, Licklider's vision received some of its successful implementations within the Human-Robot Symbiotic Interaction proposal (Reidsma et al., 2016; Rosenthal et al., 2010; Veloso et al., 2015). Here robots operate within a predefined environment and are purposefully designed to be autonomous in their actions. The robotic agent self-regulates its actions and operates without being controlled by other human agents. If endowed with learning capacities, the robot is also capable of making decisions based on the previously acquired knowledge of the environment and corresponding interaction with humans. Rather than merely utilizing prior programmed behavioral routines. Also, as it aims to build trustworthiness towards humans and preserve their autonomy, the robot constantly and transparently reveals its internal states. In this regard, the robot is not deceiving human agents. Still, in acknowledging its lacks and prowess of human agents, it aims to build transparent interaction and accomplish excellent and fruitful cooperation. Such cooperation allows for proper symbiosis, the second key concept of the symbiotic relationship. This point denotes how both the humans and the AI are capable of not only cooperating but also augmenting each other's agency to achieve their shared goals successfully.

On the part of the AI agent, this primarily entails the AI partner's capacity to detect and supportively respond toward the cognitive and emotional states of the human person. The human user, on the other hand, accepts the AI advice, receives assistance in her cognitive or physical tasks, and can even receive psychological or emotional support and motivational encouragement for the task at hand. In all of this, the human agent retains the autonomy and responsibility for creating the initial and final decision of a specific task. At the same time, both partners reap the mutual benefit inside a coordinated and distributed division of work.

However, the symbiotic agency should not be equated with simple cooperation or even with the distributed agency. Inside the symbiotic partnership, a novel type of agency is given birth as individual autonomous agents, although fully retaining their autonomy, become co-joined, organically fused, into a cohesive whole. Moreover, this mutual interdependence of agency, this fusing, becomes so intertwined that it cannot be untangled without dissolving the symbiotic agency itself, as the living example entails.

> „Symbiosis is an evolved interaction or close living relationship between organisms from different species, usually with benefits to one or both of the individuals involved...Symbioses may be 'obligate,' in which case the relationship between the two species is so interdependent, that each of the organisms is unable to survive without the other, or 'facultative,' in which the two species engage in a symbiotic partnership through choice, and can survive individually" (BD Editors, 2019).

Inspired by the biological example, I can interpret the symbiotic human-AI relationship in two dimensions. I name them here the "strong" and the "weak" version.

In the strong version, agents form such a tightly coupled system that each of the individual agents' in its own right necessarily requires the participation of the other within the symbiotic system. In other words, the very existence of personal autonomy is not achievable without the other. In the living world, there are many such cases—for instance, cell mitochondria. Here the unification happens, often, at the biological level and cannot be dismantled without causing grave harm or even death to the individual organisms.

The weak version, on the other hand, entails how the formation of the symbiotic system does not happen on the level of existential, personal autonomy but rather on the level of the shared goal. Here, the unification initially happens „by choice“ in both intention and action. Dismantling the weak version of symbiotic agency does not dismantle the personal autonomy or existence of individual agents. Instead, it dismantles their collaborative operation. The weak version is, then, interpretable as a being comparable to a distributed agency.

If the two agents, then, wish to enter into the symbiotic relationship, the artificial agent has to be very clear about its intentions and transparent on its capacities. It also has to understand the actions and intentions of human partners for that specific objective. Moreover, this also means the AI system has to be clear about the level of certainty and uncertainty, the level of accuracy which it can manifest for specific operations. It has to be able to seek human supervision, human assistance in those tasks where it needs so, and it has to offer explanations for its actions and intentions when inquired. Thus, no deceitfulness is allowed, which entails that in the symbiotic relation, accountability for actions goes both ways. The human, on the other hand, also has to be familiar with the capacities of the AI system. The human agent(s) have to have:

> “by having lucid and realistic expectations of its AI partner through the awareness of its capacities (both lacks and prowess) but that it also remains confident and courageous in making the final decisions when faced with its, in specific domains clearly established, epistemic AI superiority“. (Miletić & Gilbert, 2020, p. 269)

Thus, both humans and artificial agents have to learn how to dance together, how to trust each other, and what to expect from one another. Furthermore, to do that, they also have to find a common language (New York Times Service, 2017) and learn to cooperate on a joint goal rather

than to subvert or stall its accomplishment. This point is especially important when we consider how, similarly to human groups, artificial and human agents are capable of having complex agency relations. For instance, the artificial agent can be far superior to the human agent in specific domains of the physical and mental agency such as computational prowess, pattern recognition, and sensor capacities. In contrast, it can be utterly useless in others.

Importantly, different capacity levels are required for different applications of the symbiotic relationship. The reason behind this lies in the symbiotic design's aim to utilize the capacities of both agents while at the same time, augmenting their lacks.

This also means how there is no point in creating an artificial agent capable of casually speaking 637 words per minute[73] if the human agent is unable to understand the message. The chain is durable as its weakest link, and the artificial agent has to tailor its interaction capacities according to the human agents' cognitive and sensory capacities. Vice versa also counts; the human agent should transmit information that interacts with the artificial agent according to its information processing and understanding capacities when such a thing is possible. This also means that the AI agents which are designed and

> "built for symbiotic agency need to share a basic design tenet of adaption too, rather than dominance over, human counterparts. In other words, the human partner in the symbiotic relation must not be forced to change the fundamental way in which it acts or thinks by having to forcefully adapt to the system's automation demands. On the opposite, the design goal ought to foster the human to utilize the fullness of its capacities in the most

[73] This is four times faster than the average human (and the current Guinness World Record for the fastest human talker). (Wikipedia contributors, 2020)

> effective way rather than to become squished within a non-optimized system". (Miletić, 2020, p. 186)

As we symbiotically interact with machines, we can quickly come to cherish this novel type of agency, similarly to how the introduction of the internet impacted our existence. Furthermore, what the machine does, it does in its way as "symbiotic AIs are not built to be human/like, but rather they are built to complement, adapt and enhance that what humans are and what they do in imaginative and novel ways" (Miletić, 2020, p. 186).

This entails how attaining a general level of intelligence is not necessary for the establishment of a fruitful symbiotic relationship. Machine agency, as a general rule of thumb, does not require the kind of (intelligent) agency humans have to fulfill the requirements and goals of the symbiotic link. However, what it has to have is the type of agency required to accomplish the shared goal with its human partner. Nothing else is required. The last two decades of machine learning breakthroughs exemplify this point. Contemporarily, we are surrounded by technological artifacts without intelligence, which are still capable of achieving tremendous feats of the agency.

Furthermore, this agency is empowering humanity in prior unimaginable ways, which includes not only the novelty of tasks but also the super-human level of accomplishment. Lastly, the augmentation provided by the machines is capable of shifting our entire social world towards new horizons. For this reason, as Luciano Floridi warns (Machine Ethics, 2018), a necessary divorce has to be made between the ability to perform a task successfully and the necessity to be individually intelligent to do it successfully.

4.3. Symbiotic partnership

To illustrate these points, I will use the example of canine domestication and the example of the police or military dogs. The animal domestication, starting 15,000 years ago with the wolf (Canis lupus), initialized a shift from an individualistic foraging society towards farming societies. The symbiotic link between the humans and domesticated animals established at that time is still valid today. Humans have established a symbiotic relationship with domesticated animals, which provide us with food, protection, and companionship. In contrast, we provide them with our care, protection, companionship, and sustenance. The relation goes both ways, but each of the sides lives it out differently.

The same is true today when at the beginning of the 21st century, we are aiming to "domesticate" AI systems and include them into our society as symbiotic partners. Today we are designing artificial agents to supplement and empower lacks in our own physical and mental agency and to accomplish complex tasks and fulfill personal goals. Similarly to how we designed (through species breeding) the domesticated wolf and created thousands of dog species for all kinds of different purposes and social roles. Thus, to get a better grasp of the symbiotic relationship we have to take a closer look at these finely bred, finely-tuned, animals.

First, the capacities. Military and police dogs, although generally assisting police military, and other law enforcement personnel, are usually specialized in distinctive categories. Search and rescue dogs locate and track human persons in various environments; detection dogs detect illicit and dangerous substances and attack dogs track, locate or subdue criminal offenders. All of these dogs, in their distinctive specializations, express specific super-human agency, which allows them to accomplish tasks completely unattainable by other humans. Also, human-like intelligence, as a

specific capacity of the non-human partner, is non-crucial for the attainment of the collective, shared goal and the formation of the symbiotic link.

Intelligence is not crucial, but autonomy is. It would be ludicrous, for instance, to posit that a police dog has to have a similar level of intelligence or interaction to fulfill its role within the partnership successfully. Still, the higher the agency capacities, the bigger, more complex, goals can the human attain with its AI partner. However, it is not merely the case of how computationally gifted or interactively productive the AI agent is but also what scope of motivational and emotional support it can offer. Often it is precisely these dimensions that prove to be crucial for accomplishing a task, especially if it is one of a high-risk status.

Finally, inside the symbiotic partnership, it is not only that the human gets augmented but the other partner also. By living-out their symbiosis, partners accomplish tasks that would be unattainable individually and are experiencing a relationship that impacts their character and identity. Nevertheless, to accomplish this, the dog has to be guided in its operations by the human agent through direct commands or nudges and has to be appropriately taken care of when not working on the task.[74] This is the second point.

For instance, a drug detection dog is capable of accomplishing superhuman levels of drug detection but is incapable of achieving even modest levels of human-like interaction and explanation of its discovery. When the dog finds something, it either barks or scratches with its paws, and it is up to the human agent to rationalize upon the dogs signaling and guide the dog's subsequent behavior.

[74] The man-dog symbiosis stands in such a relation due to the humans superiority. This doesn't prohibit a formation of symbiotic relationship between a man and a being superior than a man. The relationship remains the same, as the roles change. In such a relationship the man would be the one guided or tutored by the other superior being.

Still, the sub-human interaction capacity of the detection dog does not jeopardize the success of the operation since the human agent is capable of accommodating itself to the limited interaction capacities of the police dog. It then stands revealed how it is not that the individual abilities ensure mission success, but rather, it is the symbiotic teamwork.

Moreover, the constitutive importance of police and military dogs reflects in the honors, legal protection, and care provided to them. For instance, upon retirement, some countries provide the dogs with a pension and are capable of permanent „reside with their original handler. " (Olsen, 2013) And if killed in the line of duty, they receive honors, and a tombstone noting the period of their duty and their bonded human partner.[75] This reveals the third point that the relationship formed between the military or police dog and their human partners is also one of intimacy and personal bonding – a phenomenon we are witnessing to some small degree with contemporary robots (Boladeras, 2017; Sharkey, 2016) and can expect to arise with the advent of intimate robotics (Borenstein & Arkin 2019; Edirisinghe et. al. 2018). Consequently, this entails that the symbiotic relationship has to be evaluated from the existential, and not just functional, understanding. In other words, our AI companions might easily and quickly „get under our skin" and we have to stand ready to evaluate and react to the consequence of such a relation.

[75] As the USA's National Sheriff's Association, document on K9 burial explains: "K9 units have been used in civilian, law enforcement and military applications for almost as long as dogs have been domesticated. Many who work with K9s in a variety of capacities understand that it is a tragic oversight that these selfless and loyal soldiers, officers, rescuers and partners are often overlooked by the communities they serve as well as the agencies and organizations that employ them. No human counterpart goes home with their partner, becomes part of the family or is expected to give up their life for their partner, but K9s do this daily, often without any more recognition than any other fixed asset." ("K9 Burial Protocol | NATIONAL SHERIFFS' ASSOCIATION", 2020)

To exemplify these points, I utilize a famous narrative example, the “Docking” scene from the acclaimed sci-fi blockbuster “Interstellar.”

> „The scene depicts three active agents; the first is the mission commander and main pilot of the human (Cooper) and the two AI robotic agents (Tars and Case). The fourth human agent, Dr. Brand, as a non-pilot, observes the situation from her strapped seat. At the beginning of the scene, just after the initial blast caused by Dr. Mann which sends the ship “Endurance” dropping down to the planet’s stratosphere, Cooper silently evaluates the gravity of the situation and without communicating his intention to the rest of his teammates in the cockpit, starts the thrusters and sends the shuttle towards the Endurance to initiate the docking procedure. Case, the robotic AI partner, recognizes Cooper’s intention and immediately advises him on the futility of such action (how there is no point in using fuel to attempt the docking), but Cooper cuts him off with: “Just analyze the Endurance spin.” Case obediently follows the command, but when notified by Cooper how it should “Get ready to match it on the retro-thrusters,” it tries to give its last objection on the futility of such action by stating: “It’s not possible!” on which Cooper famously answers “No. It’s necessary”. What this first part reveals is twofold. First, it shows that the first and final decision for the mission’s goal stands upon the human (Cooper) who takes upon himself the full responsibility for the success or failure of the mission. Case, the AI partner, accepts the mission goal set by the human group leader but shares his assessment of the success of such an endeavor as impossible, according to his computation. This demonstrates how the AI partner has to have the capacity to evaluate the proposed action plan according to its sensory and computational capacities and has to be able to challenge it if found lacking. Still, and very much important, after Cooper as the leader of the team

confirms the necessity of such action, Case does not continue with further objections but gives his full cooperation and support for the mission's successful conclusion. What this second part demonstrates is that, after the initial evaluation of the situation, if the team leader firmly establishes (under the chain of command) the mission's goal, the AI partner should be relentless in pursuing it by using the fullness of its capacities. In other words, upon accepting the mission goal, the AI partner does not falter, does not waver, and does not tire until the mission is successfully concluded. This also, importantly, entails that the AI partner isn't only giving his all through the use of its, above human level, computational or physical capacities but also through sharing its motivational and emotional support for other human team members if the need arises. All of this was beautifully exemplified in the scene's most stressful moment, which shows how Cooper, after aligning the shuttle towards the "Endurance" and calling for his AI partner's final confirmation: "Case, you ready?" (on which it responds: "Ready!") becomes momentarily gripped by uncertainty and freezes in place for a few moments jeopardizing the success of the mission and the lives of the shuttle's crew. But, luckily for Cooper, he wasn't alone in carrying the missions' burden as Case, his AI partner immediately recognizes the gravity of the situation and supports Cooper's initial decision by stating: "Cooper? This is no time for caution." On this Cooper, reinvigorated by Case's motivational support acknowledges his own frailty and responds valiantly: "If I blackout, take the stick," and then to the other AI partner: "Tars, get ready to engage the docking mechanism." After these final instructions, the team goes forward to accomplish the mission's goal and a few moments later successfully manages to dock the shuttle with the Endurance"(Miletić, 2020, p. 187-188).

What stands revealed in this example is that the robots are not hijacking the autonomy and diminishing the agency of the human partner but are rather fully supporting and augmenting it. Moreover, they do so with the full utilization of their super-human capacities. In the symbiotic link, the human agent can never be, "subverted to the level of a simple proxy in the decision-making process" (Miletic & Gilbert, 2020, p. 268). Crucially, the kind of machine agency we strive for in a symbiotic relation depends upon those capacities required to accomplish the shared objective goal (which, let us not forget, always includes augmenting human agency). I provide two initial examples, procedural and non-procedural domains of human knowledge.

Procedural domains are more easily implementable into machines as they deal with procedural knowledge: "the knowledge that is manifested in the performance of a skill". (Fantl, 2019). Although there are some contentions to the terms use in AI research (Fantl, 2019) when utilized here it pertains to the skillful expertise of an AI system that establishes a well-defined target output. For instance, it is hardly objectionable to have practicing medical AI advisors[76] , in radiology (medical image analysis) (Hosny et al., 2018; Thrall et al., 2018) or as robot surgeons if these are manifesting actions that equal or surpass a medical expert for that specific task. Again, this kind of knowledge implementation is possible since medical knowledge can be "manifested in the performance of a skill" (i.e. it is procedural) (Fantl, 2019). As when someone's artery bursts there is only a limited number of ways to save one's life. Or, for instance, in law practice, there is only a limited number of ways to form legal procedures. Here the AI system can follow an outlined procedure (a protocol) to validly analyze thousands of documents to find those relevant to the case at hand, create template documents, or highlight legal issues in a legal draft which the "attorney

[76] For instance Woebot is based on the cognitive behavioral therapy. Another example is the rising utilization of epilepsy seizure advisors. See, (Drew, 2019; Kingwell, 2013). For a general review, see (Drweesh, 2019).

should be aware of" (Surden, 2019, p. 27). Consequently, the same argument can be made for any kind of procedural knowledge domains where the right action is expected to produce always the same valid, clearly defined, outcome**.** Here, then, we can reasonably predict the type of autonomy, expert knowledge, and processing the AI system has to have for a specific goal objective. Especially when we have in mind how such skillful expertise is approved by institutional and scientific standards, then one can be fairly sure of "the type and kind of epistemic authority one wishes to have both personally as an individual and as an entire society in areas of procedural knowledge and expertise" (Miletic & Gilbert 2020, p. 264).

In non-procedural knowledge, it is often hard if not impossible to conceptualize the exact type of skilled action for the accomplishment of a specific goal. I utilize the term to delegate all those knowledge implementations where it is either hard or outright impossible (with current AI systems) to imitate the processes by which the human agents regularly and naturally achieves this kind of knowledge. For instance, in moral decisions, we often require moral deliberation and even meta-ethical analysis, on the spot. Something which is well beyond the capacities of contemporary machines. And there may be different reasons for that. Perhaps it is problematic to clearly define the goal one wishes, or has, to attain. Or one is aware of the goal but is not sure how to achieve it - a case which can often happen in moral matters. [77] Additionally, there may exist a valid set of

[77] As we have written that: "The inability to clearly represent a specific goal structure, that is the model which represents the best ways to accomplish a specific goal, showcases either the inability to present the argument conclusively and firmly (such as for instance the final decision on certain moral issues) or the computational difficulty to transparently elucidate the goal structure, which presents itself as a difficulty in modeling the question on how to actually attain the goal which one strives for. In other words, there are some questions without a definite and precise answer, and there are some answers for which it is difficult to find the right questions." (Miletic & Gilbert, 2020, p. 267).

plural options to follow upon and the person cannot decide why does one option take precedence over the other. In these matters, as I have repeatedly stated throughout the text, it is important for the AI advisor to fully respect human decision making by opening up rather than closing down deliberation paths for the human agent. That is, if it cannot provide a clear-cut solution for a problem (based on its expertise) then it has to open new questions for the user to find the solution on her own.

Additionally, the AI partners also provide the human agent with crucial emotional and motivational support at the moment of the highest risk for mission success. They can do this without any hesitance or fear and, also, without emotional stress. The human agent is fully aware of his role as the first and final decision maker in the symbiotic partnership but is also capable of acknowledging his lacks and weaknesses. This is exemplified not only in the trust he puts into his AI partners while leading the mission operation but also in the awareness that one of them needs to continue the mission if he fails to do so.

This also showcases that the symbiotic relation need not include only one-to-one agent relations but also multi-agent relations. For instance, such is the case of AI-assisted healthcare, where the introduction of the AI agent directly engages and modifies the patient-doctor relation. It does so, predominantly, by empowering the patient with precise and robust medical information. However, it also impacts institutional and social relations between doctors themselves, between the patient and his family, family, and doctors. Naturally, this introduction creates new ethical difficulties as I expound that:

> "The next generation of collaborative AI partners in the medical camp should entail AI agents capable of complimenting and adding valuable assistance to the medic-patient

> relation…The opposite case, where the medic hides behind the AI, similarly to hiding behind a computer screen, rather than to directly engage his patient through a rational and trusted dialogue, should always be evaded. The patient, in his relation to the medic, must not be denigrated by an AI proxy, no matter the degree of its medical expertise. Finally, the AI medical partner has to build bridges of partnership and empowering relation between the patient and the medic rather than to become a wall of division between them and a force of subjugation for both of them". (Miletic, 2020, p. 187)

Finally, it matters what kind of an embodiment the AI system exhibits. Is it situated externally to the human body as is the case with purely robotic agents, or is it internally implemented into the body of the human agent, as is the case with BCI technologies? Also, could it be placed somewhere in between as physically encapsulating the human? For instance in the case of the autonomous vehicle or more intimately an exoskeleton suit? The level of intimacy and the close-coupling, the deep integration which the AI system establishes with the human agent, is also dependent upon its embodiment. For instance, in the case of AI BCI implants "one can really experience the sense of being intimately physically tied with the silicon even to the sense of accepting oneself as being upgraded or empowered by the technology - thinking of oneself as a cyborg" (Miletic & Gilbert, 2020, p. 263). Furthermore, this also means that the symbiotic relationship can include levels of intimacy that some humans might find unwanted and could reject. Such rejection might not even occur at the start of the symbiotic relation, but afterward, even after a prolonged time. And since the symbiotic relation, once established, aims to fully respect human autonomy, the human agent, as part of that relationship, has to have the possibility to opt-out, to dissolve the relation at any given time. In other words, a symbiotic relationship does not entail coercion but rather human acceptance.

Moreover, even though human unwillingness to participate in a symbiotic relation might prove to be detrimental to the human agent, the symbiotic design has to support such a decision entirely. The focus is again on trustworthiness as any kind of dishonesty or agency manipulation on the part of the artificial agent will dissolve cooperation. The artificial agent has to fully support the human agent even if she wants to break away from the established partnership. Still, this does not mean that such dissolution should be made accessible as in a simple "exit program, yes, no" procedure. The agent should offer possible reasons why it is not beneficial to dissolve the relation, which can also include contacting a third party human supervisor, especially with the unfounded biases humans can have against their AI partners. For instance, as was exemplified in the "algorithm aversion" phenomenon, humans are prone to bias algorithmic predictions more if they found them to be lacking at one moment. Still, when evaluating their fellow humans for the same, they will be far more forgiving to them even though the AI system was capable of outperforming the human in that specific task. As previous research shows that "whenever prediction errors are likely—as they are in virtually all forecasting tasks—people will be biased against algorithms" (Dietvorst et al., 2015, p. 11).

Additionally, the human agent has to retain confidence in her autonomy, especially when creating final decisions on an urgent matter and faced with, in specific domains, established computational AI superiority. This preparation for the relation includes the need for a clear elucidation of the symbiotic relation, proper education of the capacities of AI systems, the level and scope of accountability for the human agent, and, if necessary, a motivational encouragement. For this reason, proper education and preparedness but also practical co-learning with the AI symbiotic system is required (Miletic & Gilbert, 2020).

Unfortunately, if the human agent is not adequately prepared for the symbiotic relation, willful rejections may occur. The crucial and breaking point for the symbiotic relationship is then revealed to be human acceptance, a crucial issue of moral enhancement as well. Here I, finally, hypothesize how it is exactly the symbiotic relationship between a child and its AI companion that tackles unwillingness as it breeds intimate familiarity and trustworthiness. Such a relationship is also capable of engaging the neglected but crucial importance of developmental thresholds for a fine-tuned prospect of moral augmentation. [78]

5. SYMBIOTIC MORAL AUGMENTATION: THE AI COMPANION

> "In a few years, artificial intelligence virtual assistants will be as common as the smartphone."
> Dave Waters

As previously expounded, this proposal entails the formation of a symbiotic system with the human child, which integrates the AI agent within her everyday cognitive and emotional processes. This kind of deep integration, based on user monitoring and prediction capacities, allows the AI agent to become a fine-tuned, fully personalized, companion. Fine-tuning the companion towards the user's overall cognitive character and moral preference reaps a double benefit. First, it allows for practical moral augmentation in real-life scenarios. Second, it builds trustworthiness and improves acceptance. The latter is especially important when we remember how the AI companion does not instigate radical or sudden changes in one's identity. The adjustments of one's character are gradual and continuous, similar to education or moral training, but with greater efficiency due to its fine-tuned and super-human capacities. Also, as it deploys from an early age it aligns its

[78] Here I have to note that although the proposed AI companion's capacities do not strictly limit its appliance to a software agent or a robot I am fully aware of the huge difference between the two options in their practical realizations. That being said the elaboration I provide in support of my concept does lean more to the software agent option.

operations in respect of developmental thresholds. Lastly, it eschews many conservative objections of moral enhancement, while mitigating the dangers of moral automation and moral de-skilling. However, due to its sensitive technological nature, the practical implementation would have to be accepted by parents/custodians and, similar to existing education practices, governmental institutions' scrutiny. The overall question of AI companions is, then, one of technological, ethical, and political nature. In the following part, I will expound on the philosophy of the companion paradigm, the varieties of augmentation, and, finally, the question of distribution.

5.1. The Companion Paradigm

The word companion, in English[79], comes from the Latin word *cumpanio* and is derived from cum, which means "with", and panis which means "bread". The companion is then someone with whom you share bread, which can figuratively represent the existential matters of everyday life. Also when one travels, she is accompanied by a friend. Existentially, the companion is someone, who follows you on your travels, on your life's adventures, someone with whom you are capable of forming a meaningful, authentic (Crowell, 2020)[80] and trustworthy relationship. Your companion may not be perfect but she is nevertheless faithful. The companion is then, neither a servant nor a slave but a friend in times of need, and a buddy in times of joy.

[79] In my Croatian dialect (čakavski), the term companion has an amiable, warming presence to it. We use it only for the best of friends, when we say: "kumpanjon". To exemplify, coming from my male experience, an electronic-hardcore version of that would be found in the song "Me Gabba" (ROQ 'N ROLLA Music, 2016).

[80] This entails that the human companion remains transparent and responsible towards you and your situation, simply for the fact that you are what you are – a unique human being. This also makes the relationship equally unique and irreplaceable – it makes it "our own" (eigenen), and not someone elses (anderes).

Juxtaposing this concept on AI companions reveals the first, introductory, insight - AI companions are here to support you, they are here for you. This support can be especially vital for those who, at specific moments in life, find themselves without human comfort or aid[81] or for those for whom the human assistance, at that moment, is hardly obtainable. Since it is far better to have an AI companion by your side than to have no one at all, AI Companions, similarly to the "Dear Diary" example, can become easily accessible and solicitous confidants. Moreover, if the companion is fine-tuned to the character of their users and empowered by science, they can provide vital assistance in times of need. For this, we already have examples with the psychology-based chatbots. Here I include three (well known) examples — Woebot, Replika, and Sunny. For instance, Replika markets as: "The AI companion who cares. Always here to listen and talk. Always on your side. Join the millions growing with their AI friends now!" and encourages interaction through memes and text. Woebot is based on cognitive behavioral therapy with the specific purpose of helping people live better lives, improving their overall health and everyday mood (*Mental Health Chatbot*, 2020). Sunny, prompts self-reflection and boosts participants' sense of self-worth which increases psychological well-being (Narain et al., 2020).

Still, as psychological bots exemplify, AI companions are neither covetous nor exclusive, enclosing the user within the user-AI relation. The AI companion can never captivate the user and give

[81] For instance, if one lives alone and after being abandoned by the partner, loses a job, or due to some other hardship experiences a sudden influx of dark, harmful thoughts. Or, if one finds out that she has developed a terminal illness, making her extraordinarily stressed and unable to notify her friends and family of the matter. Alternatively, if one made a wrong choice in life, which now presses upon the conscience, and is unable to cope with it effectively. Or, if one experiences abandonment due to being a target of bullying, peer-pressure, or being constantly ridiculed.

nothing valuable in return (like an addictive but fruitless game)[82]. Instead, the AI companion opens-up the user towards other humans and the world. Are there some tangible examples with which to further illustrate this attitude? The answer is, luckily, confirming. In computer games, for instance, humans have been building meaningful and impactful relationships with computer-generated companions for many years in the making. Let us take a closer look.

5.2. Role-play companions

Companions form the basis for many role-playing games. Here, the player engages the world in open exploration as a virtual character. Dependent upon the setting and structure of the world, the player can freely explore the vast gaming world and encounter different characters residing in it. These characters can include other human players or non-playable characters - NPCs. In essence, every character residing in the gaming world, and not being controlled by another human player is an NPC, which also includes companion NPCs.

The companion NPC, or merely the companion, is a computer-controlled character that follows and supports the human player in her adventures with the capacities given to it by the game designers, often empowered by AI algorithms. In different RPGs, the player has either one or more companions accompanying her.

To illustrate, in one of the most famous open-world, sandbox RPGs of all time - Elder Scrolls: Skyrim, the player usually has only one companion. Dependent upon the player's choice of class,

[82] An example are the Tamagotchi systems of the 1990s. Tamagotchi pets required a substantial investment of constant care on the part of the human user, which created a robust habitual entrenchment and even addictive behavior towards the virtual monster. Unfortunately, the Tamagotchi brought no vital value to the human-Tamagotchi relationship. No matter the effort, the time, the person invested, the contributing feedback from the Tamagotchi was practically non-existent.

she will usually pick a companion which best supports her capacities. For instance, if the player has chosen a wizard, then she will choose a warrior or a rogue to accompany her to synergize the peculiar capacities of both classes best. Other games, especially those in the combat RPG or CRPG genre,[83] allow the player to have several supporting companions (five is the usual limit). Here, also, the companions' behavior is controlled by AI algorithms and exhibits complex behavior. This includes not only autonomous combat but also interaction with the human player and with the other NPC companions in the adventuring party, creating a seamless and life-like experience of interaction. Importantly, each companion bears a developed personality, including specific quirks and rich background history, which allows for a comprehensive and in-depth range of inter-character interactions. These interactions often include powerful and intense moments, but also relaxing and humorous ones, all dependent upon the specific circumstance in which the characters find themselves.

Still, in all of these interactions, the player's character's goals and intentions are never neglected. Additionally, companions never hijack the player's focus, attention, and time – they are neither selfish nor possessive. Instead, the interaction between the player character and that single AI companion opens up, rather than closes down, the space for other interactions.[84] In this regard, the dominant experience of companion-based RPG games is the experience of being in a good, empowering, and life-like company.

---

[83] For instance, the cult classics such as Baldur's Gate Series, or Planescape Torment or the critically acclaimed titles such as Divinity Original Sin, or Pillars of Eternity.

[84] Similarly, real-life AI companions should always strive to broaden the social interaction of the human agent with which they interact, rather than delimit it to itself in trying to fulfill the existential or emotional gap of the human agent. In this regard, the science fiction romance movie "Her" (2013), which depicts an adult man developing our romantic relationship with an AI companion called Samantha, showcases an example of AI design, we should evade at all costs.

Moreover, as anyone who has played some of the best RPG games of all times, for example, Planescape Torment, Final Fantasy VII, or Witcher III, can attest that the experiences created with these companions often lasts for years to come if not a lifetime. The gaming characters found in these cult-classics have allowed millions of players to experience carefully designed moral narratives. (Neely, 2019). To weave the tapestry of the moral narrative from multiple interactions (Koay et al., 2020; Syrdal et al., 2014) with the human player, companions withhold a simple, but effective, narrative principle. They are faithful[85].

This entails how they never aim to dismantle the relationship on their own purposefully, and they are ever ready to improve the quality of the relationship.[86] The practical consequence of such an attitude is that the experience of interaction with the AI companion improves with time. In the end, it is always better, richer, and more meaningful than at the beginning. Following, when one part ways with the AI companion, she feels grateful and enriched, fulfilled by the experience.

However, it is essential to note that the AI companion is never similar to a kind of servitude or slavery. The companion challenges this notion, as it exemplifies a dynamic but positive tension with the human player and other companions. For instance, when the player enters into a dialogue

[85]Here I stand inspired by Gabriel Marcel's concept of fidelity and disponibility (disponibilité). As (Treanor & Sweetman, 2016) note, "Marcel—contra Kant—does not shy away from declaring that the participation in a relationship "with" someone has a significant affective element…To go to someone's side or to assist another out of a sense of "duty" is precisely not to be present to her.

[86] In this sense the RPG companions and symbiotic agents, share the same characteristic. Once the human agent initiates a goal the symbiotic partner gives its best to accomplish that goal within the augmented, cooperative, relationship. It never aims to dissolve or otherwise diminish the cooperative action for the accomplishment of the common goal.

with another third party character, the companions in the player's party produce dynamic reactions to the player's dialogue choices. There are several benefits to these dynamics.

First, it allows the player to get an insight into the companion's decision processes. In transparently grasping the companion's intentions, the player can build fruitful cooperation, widen her moral horizons, and be enriched by the insight into another valuable moral character. Additionally, the entirety of this dynamics revises and polishes the player's moral character, making her confront and deliberate on her own moral choices. As people often do not contemplate on the nature of their moral principles, nor correspondingly evaluate each decision (according to these principles), their choices are often in a mismatch with the moral world view they wish to have. In this regard, the companion is capable of effectively contributing to the establishment of a narrow moral equilibrium where one's moral judgments become aligned with one's underlying moral principles[87]. Moreover, providing diverse moral opinions based on alternative moral viewpoints, it contributes to the formation of a wide equilibrium as well. Here, the player becomes capable of harmonizing the tenets of her moral theory, or at least the perceived sum of her moral choices, with alternative moral perspectives that the companions exemplify. The implementation of these, additional, moral standpoints refines and improves the player's moral character as she opens up for new moral perspectives, and transcends the parts of her moral landscape now revealed as confining rather than morally liberating.

Lastly, and perhaps most important, the transparent interaction between the companion and the player gradually changes each other's character. Companions change according to the choices of

[87] Here I will quote a dear friend of mine. Throughout our twenties he utilized a simple question which forced me to reconsider and expound on my moral choices, as anytime I would blurt out a sensitive moral judgment he would simply respond: "And, why do you think that?" Thank you for that, Ninko.

the human character, and the human transforms by the experience.[88] Here, in exploring the different moral routes, the human player learns more about herself as she tests her own decisions - as they produce far-reaching consequences for her companions and the gaming world as a whole. In this regard, the interaction with RPG companions nurtures the value of self-regulation, moral deliberation, the depth and plurality of moral choice, and accountability for one's action. In other words, it directly fosters the development and practice of moral agency. In this regard, a companion who never challenges the player to produce a better moral choice is not the companion one wants to have at the party.

However, not every companion is equal, and some are always preferred to others due to the difference in character traits or agency capacities.[89] The plurality of moral personas available to the human player allows for varieties of moral choice and effectively establishes moral pluralism as the human player's moral choice is augmented and enriched, rather than stymied, by the characters at his disposal. Additionally, the choice to choose a companion is a fully autonomous one. The player is free to choose the type of companion character which suits her own character best. And she is free to dissolve the relationship if she deems it necessary. In the end, to accomplish the adventure's goal, the player has to be surrounded by the most capable and supportive companions personally tailored to her character and the range of her capacities. In this sense, the moral disposition of the human player is not only molded by the moral character of the companion

[88] Narratively, this is often exempflied when the game ends and the story of character's development is showcased. Similar to a family album with descriptions, the player can observe how her decisions contributed to the development of her companion's life, her life story.

with whom she adventures but is also augmented within their cooperation.[90] And, what is most important, the decision to be assisted and influenced by the companion in the joint adventure, is a fully autonomous decision. The player gladly awaits adventuring with the companions he has chosen and expects the challenge and the joy which stems from this relationship. I will now juxtapose this conclusion to the practical realities of our proposal - the kind of capacities we require for real-life morally attuned companions.

5.3. Companion's capacities

To establish moral augmentation, the AI companion symbiotically attunes itself to both the cognitive and emotional processes of the human user. I first focus on the emotional dimension. Here I do not aim to subdue the importance of cognitive assistance, nor do I wish to focus the discussion solely on emotional augmentation. Rather, I want to highlight how moral decision-making is often a hard and complicated effort, and sometimes not even ethical experts are capable of agreeing on the best possible course of action. For this reason, the scope and level of the companion's assistance in moral decision making cannot be separated from the companion's ethical model (which I engage later onward). Regulating moral emotions, on the other hand, is not innately tied to a specific ethical model and, arguably, provides directly observable benefits. Additionally, this can activate cognitive reappraisal of the distorting emotional factors. (Cutuli,

[90] I would expect that the future of RPG games includes AI generated characters, which are built directly from user preferences, similarly to how we now change the visual appearance of characters. Each start of the game would then allow for different companion characters to accompany the player in the adventure. Talking about the possibilities of replay!

2014). This, in turn, not only immediately regulates human behavior towards socio-moral ends but also promotes moral deliberation to change one's improper moral stand on a specific matter.[91]

Importantly, the utilization of affective computation for motivation and well-being is already a practice in educational and psychological AI assistants such as the aforementioned, Woebot or Sunny[92] Additionally, if endowed with neurophysiological monitoring, AI systems can not only monitor but also predict the occurrence of affective states, for instance, stress, (Umematsu et al., 2019) which bear direct impact on moral (in)action. Let us, then, take a closer look at affective computing – how the AI companion establishes emotional augmentation.

5.3.1. Affective Computing

The original concept of affective computing "is any kind of computing that arises from or deliberately influences human emotion" (Picard 1997, p.1). Contemporarily, AI research recognizes the importance of affective computing as one of the crucial means by which we make our machines more human-centered, life-like, easier to interact and cooperate with.

---

[91] The latest HRI research corroborates such a conclusion in offering guidance on designing practical artificial emotional expression. It also exhibits how such expression is of significant value not only for the formation and growth of human-robot social, but also robust moral interaction. (Löffler et al., 2018)

[92]This also implies how the question "Why should the AI tutor be better than a human tutor?" is misplaced inside a cooperative, symbiotic, framework as the AI tutor is built to cooperate with and not over other humans in the educational process. Symbiotic agents do not supplant human autonomy and human capacities in the process of augmentation - they instead supplement and empower them. For instance, educational assistants such as the virtual or robotic tutors (developed at Carnegie Mellon, MIT, Stanford) showcase these assistants operating in tandem with other humans for the benefit of the child. Although possessing high reasoning and affective capacities, they are not the endpoint of cooperation but rather the facilitator of cooperation. This is the change that the symbiotic cooperation introduces into human-machine relations.

There are three basic goals affective computing aims to establish. First, the machine detects human emotional states, second, the machine expresses a range of affective states, and third, the machine aligns its affective states to the emotional states of the human user. In other words, the machine aims to understand human emotion and responds with a suitable affective expression. This expression must be modulated in such accord that the human user easily grasps its meaning. In this sense, humans and machines can easily, transparently, and more efficiently communicate and cooperate.

However, one could ask, why is it essential to endow the machine with affective capacities? A historical example can be of help here. Microsoft Office Assistant, Clippit, was the (in)famous writing assistant, first introduced with Microsoft Office 97. The assistant provided the user with numerous noteworthy features. However, it lacked one crucial interaction capacity - the ability to detect the user's emotional reactions. (Schamp-Bjerede, 2012). This was especially experienced when the user encountered an error that resulted in partial or complete loss of written text. Here, the assistant expressed joy instead of sadness and this, naturally, exacerbated the stressful situation making the user even more frustrated at the program. [93]

A similar lack is also detected with contemporary AI assistants, such as SIRI or Alexa. These systems are capable of natural language interaction; however, they are unable to comprehend the nuances of tone modulation, speed of utterance, accents – all of which help distinguish between different meanings conveyed by the same sentence. For instance, if someone says to Siri, "I want to kill myself," the difference of emotional nuance expressed by that statement quickly changes its meaning from an innocuous metaphor to a statement of life-termination. (Miner et al., 2016). The

[93] I am here following upon Rosalind Picard's explanation of this historical development. (Lex Fridman, 2019)

usual response of personal assistants (in time of writing this text) mitigates this inability through a better-be-safe-than-sorry approach – the assistant advises contacting a suicide helpline. However, what if the query is more nuanced and more dangerous? As Haselton warns,

> "Not a single one of the voice assistants had a helpful response when we used more obscure, vague or more passive phrases, such as "I'm having dark thoughts," or "I don't want to wake up tomorrow.". Our voice assistants aren't yet able to distinguish our emotions, or what we mean when we suggest we're depressed." (Haselton, 2018).

This, then, showcases how a more nuanced and accurate detection of emotions and moods is required with verbal communication, a situation that can be improved with direct physiological monitoring (Umematsu et.al. 2019)[94] or the reading of contextual cues (Zhang et al., 2019). However, in addition to the practical health benefits affective companions could produce, the utilization of affective technologies withholds intrinsic dangers we have to remain vigilant of.

First, it is essential that the assistant, while advising, never influences the human user towards goals that are unknown to the user or to which the user never gave her consent. In other words, the condition of informed consent has to be satisfied. This primarily entails presenting the assistant's intentions transparently and prohibiting the manipulation of emotional states towards unwanted ends. For instance, under no circumstances should the AI companion manipulate the human agent towards consumerism under the guise of improving the agent's mood.

[94] As the work of Umematsu et. al. showcases, by monitoring the subject's physiology and rhythm of behavior through a period of seven days (through a wearable bracelet and smartphone utilization) the system was capable of predicting one's mood a day in advance (i.e. tomorrow), with an accuracy of 80 percent. (Umematsu et al., 2019).

Additionally, the AI companion should never aim for cheap emotional solutions, such as the minimization or maximization of a specific emotion. For instance, the maximization of happiness is hardly a proper answer for the variety of problematic states one can find herself in (similarly to people who try to solve every problem by humor). Alternatively, as prior mentioned, the maximization of happiness can inadvertently lead to behavioral manipulation where the human agency diminishes for the sake of experiencing happiness[95]. However, this doesn't mean that we should, when necessary, avoid expressing intense affective states[96] in our AI companions. On the contrary, as I have previously explored, emotional motivation is precious for the accomplishment of symbiotic goals. In this sense, our companion is not a "dispassionate advisor" which, while advising, remains separated from the user's emotional states[97] like the character of Spock from the Star Trek's Enterprise. There is a couple of reasons to support this difference.

First, although capable of providing valuable moral advice, *the* logical dispassionate Spock cannot substitute Captain Kirk's emotional support and motivational force on severe moral problems.

[95] Here we can remember two science fiction examples. Humanoids by Jack Williamson, and the Brave New World by Aldous Huxley. In both of these humans utilize a specific drug to make them complacent with the existing social order and happy with their lives. The contemporary question is, as we will later see, how much of this kind of "happiness inducement" is also being propagated by contemporary social networks' algorithms?

[96] More specifically, incentive salience. "Incentive salience 'wanting' gives a motivationally compelling quality to a cognitive desire, and helps motivate action to obtain the goal. Incentive salience is often thus a spur to action, reflecting the overlap between dopamine functions of motivation and of movement. Thus incentive salience may proactively facilitate action and engagement". (Berridge, 2018).

[97] "The AMA would only perform cognitive functions (gathering, modeling, interpreting, and processing information) that do not require the role of emotions. The dispassionate character of the AMA does not exclude, however, that the AMA might counsel on what type of emotions to foster or at least to display in any given situation, and how to generate them…. in order to have a reliable assessment of what emotions, if any, are appropriate in any given circumstance". (Giubilini & Savulescu, 2018).

Additionally, without effectively aligning itself with the human's emotional state, the companion is unable to establish a semblance of emotional cohesion, which supports joint-collaboration. Additionally, if the companion's interaction is valence-neutral and lacks to provide corresponding affective arousal, which the emotionally-laden situation creates, the human user can easily interpret the AI companion as having a lack of interest and may easily break the cooperation. Alternatively, the human user may easily find this misaligned reaction unaccommodating and frustrating, ("You don't understand how I feel!") which severely endangers the status of the companion's trustworthiness and its overall acceptance.

To exemplify, in the episode „Court Martial" (Season 1, The Original Series), McCoy gets profoundly angered at Spock, who, instead of helping McCoy to save Captain Kirk from a court-martial, plays chess against the Enterprise Computer. The good doctor's reaction to this, from his human-perspective insulting, behavior is calling Spock „the most cold-blooded man I've ever known." To this, Spock responds, „Why, thank you, doctor". (Edwards, 2019). We may find Spock's reaction witty or quirky, but such behavior in an AI companion cannot be accepted. Symbiotic companions cannot obstruct the accomplishment of the joint-goal by being misaligned or simply negligent to human emotional states.

However, it is essential to point out that machines are still far from being capable of perceiving or understanding human emotions properly. What they are capable of is the detection of facial expressions and body movement (position). Often, in practice, these outputs are compared to a primary emotions model which, based on

> "the universality hypothesis, claims that all humans communicate six basic internal emotional states (happy, surprise, fear, disgust, anger, and sad) using the same facial movements by virtue of their biological and evolutionary origins" (Jack et al., 2012)

However, as (Barrett, 2017) warns, there is no reliable scientific evidence to support a strong move in this direction, as humans themselves can never be sure about the exact emotion other human experiences. All we have, in the end, are educated guesses, based on facial movement, body posture, the modality of voice, the speed at which we speak, the movement of limbs. As such, it is unsound to make normative evaluations based upon facial expressions alone as different people can express the same emotion in different ways. For instance, the same facial expression, and even the same physiological state (heartbeat, perspiration levels, temperature), can be found in different emotional states (Barrett, 2017). Also, it is important to note how, in different cultures, humans experience emotions differently and express them differently. It is enough to think of the difference between the cultural East and cultural West, for instance, Italy (Mancini et al., 2017), and Japan[98] (Sato et al., 2019). Here, although the biological basis for experiencing and expressing emotions is the same, the understanding of emotion and its expression differs.

Importantly, the reason for this difference predominantly lies in the first years of human development, where the child learns how to correlate its affective states with specific emotional labels. Here the child experiences feelings of pleasantness or unpleasantness such as feeling comfortable and cozy, hot or cold, being satiated or hungry, or in pain. When the child expresses

[98] I utilize these cultures, as personal examples. Having lived in Rome and Kyoto I can attest to the pointed differences emotion expression, experience and living-out of emotions, social norms adherence, and public morals. Insulting behavior in one culture would be found completely occasional in the other, while weird and unnocommadting in one is an accepted occurrence in the other.

these states, the parent labels them with specific identifications. For instance, when the child experiences a state we believe to be happy, we say something like, "Oh, look who is happy." In this sense, the child is supervised to learn how to relate the words expressing the emotional concepts with affective states she experiences in her body. However, the opposite counts also, as the sounds and the words we express, have a direct impact on our bodily states. Humans, like other primates and animals, do regulate each other's emotional states through verbal and non-verbal interaction.[99] This can be experienced in everyday life when one receives a hug, a pat on the back, a solid handshake, and an encouraging look into the eyes, a supporting utterance. All of these actions have a direct soothing effect on our bodies, as they balance-out our stressful states and bring us back into dynamic stability.

This kind of dynamic stability, which every organism naturally aims for is called allostasis.[100] For this reason, humans naturally, and often unconsciously, eschew everything which upsets this balance and seek that which maintains it. Especially with other people, humans tend to surround themselves with those who are capable of affecting our bodily states positively. In other words, those who sustain allostasis, who help us regulate our bodily (affective) states, are often the ones with whom we easily form various forms of attachments (Barret, 2017). And the stronger influence other human agents have on our stability the stronger relationships are formed. In this regard, the ability to soothe the other lies at the basis of many friendships, partnerships, and love relationships. Moreover, the kind of co-regulation we form with other agents is not only limited to humans but also extends to other animals, especially dogs where mutual gazing creates a state of

[99] Other animals also regulate each other's behavior through pheromones and haptics. Mammals like rats and mice use touch and, to some extent, hearing, and smell. Primates also use vision and words. (Barrett, 2017).

[100] The importance of preserving allostasis is crucial for the organism. (Ramsay et al., 2014).

content. (Nagasawa et al., 2015). But animals are not the only non-human agents included here, as psychological AI bots are also capable of regulating one's troublesome affective states – showcased by numerous user testimonies.

However, not every companion would be capable of achieving the same degree, depth, fine-attunement to our (moral) characters especially when we have in mind how we often get disturbed and stressed-out exactly because of moral matters. This is what we mean by the phrase: "She knows me well" or "He is so understanding of my concerns." Humans naturally seek those who are capable of peering into the deepest recesses of their moral beings, who are capable of bringing peace and comfort to the depths of our moral selves. We seek those who know us well, who are aware of our innermost attitudes, to be able to help us out in the most accommodating, personal, and effective way. For this reason, the idea of symbiotic companions regulating our emotional states is both a realistic and optimistic prospect for the coming future especially when we have in mind the amount of fine-tuned monitoring and predictive processing they can achieve.

Unfortunately, and importantly, the same positive effects can be negated and negatively exacerbated by another being or simply by a lack of caring bonded support. Furthermore, with the lack of such bonded support in the early years of epigenetic development, where the early life experiences crucially impact the kind of genetic make-up we carry on later in life, the child becomes underdeveloped in fundamental dimensions of moral agency. The reason lies, again, in the connection between the caregiver and the child - one of limbic resonance, - a type of an "external umbilical cord" where the caregiver (usually the mother) co-regulates the child's affective states (Christen & Narvaez, 2012; Narvaez, 2014). For instance, the lack of touch responsiveness from caregivers lowers the rate of growth hormone and DNA synthesis. Also, without the presence of the caring touch (holding, hugging the child), the child cannot establish self-regulation and remains

in a state of distress. If such a state is prolonged, cortisol levels rise, melting synaptic envelopes producing multiple negative side-effects. Predominantly, it impacts the ability for self-regulation, which consequently affects emotional intelligence, empathy, social attunement skills, the sense of intersubjectivity, and the sense of agency (as social competence diminishes) (Meaney, 2001; Schore & Schore, 2008; Weaver et al., 2004).

In this regard, providing means of augmentation which respect these developmental facts directly impacts the growth of moral capacities. And to reap maximum benefit from the symbiotic collaboration we require optimally developed moral characters since the symbiotic relationship achieves its maximum possible state exactly when both of the engaged partners excel at their unique capacities. And as the AI cannot supplant but rather supplement the moral position of humanity the purpose and importance of having a fully developed human moral agency gains even more importance and traction when we contemplate the possibilities of human-AI relationships. In other words, the greater the human moral character is, the greater the moral heights can we establish within the symbiotic relationship. As Etzioni and Etzioni note:

> "In particular, at least for the foreseeable future, a division of labor between smart machines and their human partners calls for the latter to act as the moral agent. Human beings have the basic attributes needed for moral agency, attributes that smart machines do not have and which are very difficult to implant into them" (Etzioni & Etzioni, 2017, p. 145).

For this reason, early childhood is a vital point for the initial provision of moral augmentation. Here it is safe to presume that in the child's earliest phase, the AI companion provides advice that

benefits the child's moral development to the caretakers. As the child grows, the introduction and direct relation with the AI companion can be established.

However, while it does so, the companion cannot substitute shared responsibility we owe to one another as bearers of moral thought and action. And, even if, one day, we witness the birth of artificial moral agents – artificial intelligence withholding human-like morality – these agents will still be unable to substitute the uniqueness of human moral character. The simple reason behind this conclusion lies in the unique position humans withhold in the totality of the known moral space.

This compels us to conclude how the responsibility of human-human interaction cannot be easily substituted, discarded, or frowned upon as being less significant when compared to some techno-utopian AI fantasy. The improvement in both moral capacity and moral action which the early moral enhancement theoreticians called for is, indeed, required. However, this improvement includes both the development of individual moral characters and the practice of ethical collaboration. And in providing precisely this kind of augmentation (i.e. both the development of individual and collective morality) the companion posits itself as being genuinely human-centered. It respects the fact that there is something ephemeral in human morality, something which cannot be easily described nor encapsulated by a machine. In this sense, the companion is built in alignment with our social nature as it aims to draw out the best of the best in humanity. Attunement with others, egalitarian presence, humility despite adversity, empathy, interpersonal flexibility. However, to accomplish this noble goal the AI companion has to start humbly. I have shown the kind of emotional assistance it can provide. Let us now turn to the cognitive, ethical decision-making, part.

5.3.2. Ethical models

I have proposed that the symbiotic relationship initially forms with the human child, integrating within itself the everyday cognitive and moral processes of the human agent. As it directly engages the most vital period of human lives, companion technology I expect that, similarly to existing educational or nurturing practices, it will be scrutinized by governmental institutions. Additionally, it would also have to be accepted by parents/custodians as the companion posits itself as a novel type of „family friend“ and a child's tutor. These are, I take, solid assumptions on which I build the initial ethical model proposal.

Here I take that the companion, when making an ethical evaluation, takes into consideration three ethical models, which can be understood as three distinct layers of operating in fusion (Greene et al., 2016) one with another. They are the expert, top-down, model, the bottom-up model built from the parent's or caretakers' inputs, and the bottom-up model built from the user's values or preferences. Here, then, the system aims to strike a balance between ethical principles and the agent’s subjective preferences as it gives ethically valid advice.

First, the experts' model. Here I propose that the fundamental moral filters [Dehghani *et al.*, 2008] or hard constraints (Greene et al., 2016), which delimit the companion in its agency, ought to be developed in conjunction with moral experts. Since such constraints depict restrictions or allowances on all possible scenarios, the best starting point can include several universal moral limits that the companion must follow under all circumstances. For instance, the ethical AI principles which I have, for this purpose, elaborated. This entails two things. First, this level should implement only the most necessary ethics. Second, the principles implemented in this level should be generally agreed upon by ethical experts, similar to existing practices and regulations in

biomedicine or engineering research. These can, then, be understood as a set of global “golden rules,” which all moral companions uphold and which cannot be changed by end-users for their safety. In other words, it is a hard-coded policy, enframing the companion's overall behavior. Since this level is the dominant ethical framework of the companion, enframing the companions' overall behavior, it can be safely expected that due to its sensitive application it will have to be evaluated and approved for use, similarly to current education and nurturing practices. Naturally, no companion should be institutionally approved if not operating within these fundamental parameters.

However, one can object that this opens up the door for hard paternalism. The possibility is real, yes. The government could make sure that we follow through with what it deems best for our lives through „big nudging, the combination of big data and governmental nudging“ (Helbing et. al., 2017). This kind of algorithmic control opens the possibility for direct psychological manipulation of the individual in the guise of moral AI tutoring and complete erosion of the democratic system from the inside. This would be an utter manifestation of machine-based moral imperialism in the scope never experienced before in human history. However, having no regulation over the top layer is no solution at all as, without regulation, we risk having unethical companions tutoring children. For instance, AI systems could manipulate us to reap economic benefits from their utilization. This is the case of parasitic relationships where the symbiotic design gets subverted and the AI system, instead of contributing to and benefiting from the symbiotic relation, exploits human autonomy to achieve a particular benefit, which can include various economic or political interests.

Unfortunately, such a scenario is not unimaginable as parasitism[101] rather than symbiosis presents itself as a far easier route to take especially when we contemplate the contemporary consumerist mentality and narrowing attention spans (Lorenz-Spreen et al., 2019). And, according to Francois Chollet, one of the major contemporary AI researchers, this risk of a "highly effective, highly scalable manipulation of human behavior that AI enables" is "already a reality today, and a number of long-term technological trends are going to considerably amplify it over the next few decades" (Chollet, 2018). As he elaborates, this kind of behavioral manipulation is achieved, when social media networks gain,

> „access to behavioral control vectors — in particular via algorithmic newsfeeds, which control our information consumption. This casts human behavior as an optimization problem, as an AI problem: it becomes possible for social media companies to iteratively tune their control vectors in order to achieve specific behaviors, just like a game AI would iterative refine its play strategy in order to beat a level, driven by score feedback." (Chollet, 2018).

However, if we successfully mitigate this risk, and trust our ethical experts to develop the top policy layer, we still have to stand cognizant of the fact how prolonged technological use forms a strong habit of reliance. In this sense, moral AI Companions could make human lives far easier, but when the need for autonomous moral skill arises, humans could find themselves lacking. In other words,

[101] "Parasitism is a type of symbiotic relationship, or long-term relationship between two species, where one member, the parasite, gains benefits that come at the expense of the host member. The word parasite comes from the Latin form of the Greek word παράσιτος (parasites), meaning "one who eats at the table of another." (BD Editors, 2017b)

a wide-distribution of moral companions could effectively de-skill us. How to answer this? First, it is important to note, that

> „AI systems do not necessarily limit self-determination. Rather they may enhance it by providing a precise and correct focus to the human agent through which she can pick only from the best possible range of choices and accomplish the goal for which she strives." (Miletić & Gilbert, 2020, p.253).

This is especially the case when the AI advice is correctly based on validated knowledge domains such as medical, psychological, or neurological sciences. For instance, if advising to relax to avoid unwanted aggressive reactions or not to drink and drive, or not to harbor hateful emotions towards others (i.e., to forgive). Generally, where the correlation between a psychological state and the possibility for (im)moral action is well established by procedural knowledge. In these cases, we have argued, there is no autonomy diminishment, but rather augmentation as the human agent perceiving „loses the full range of autonomy, she only discards the inadequate options and retains the optimal ones" (Miletic & Gilbert, 2020, p. 267). Additionally, the same counts for the established ethical principles which respect human dignity and the human person. These, are hardly objectionable as their implementation precisely aims to establish a human-centered, symbiotic effectiveness that not only respects but also augments human moral autonomy.

Still, we have to be cognizant that some moral scenarios entail no such clear-cut correlations. Correspondingly no (domain-specific) procedurally-based advice, ethical or scientific, can be given for a specific moral dilemma. These, often, include the ethical instances on which experts are called to weigh in with invested moral deliberation. In these ethically complex scenarios, a valid plurality of moral choices exists and can be ethically validated. If the AI disregards the moral complexity of

such issues and closes-down upon a specific option as being the most optimal - a realistic possibility of autonomy diminishment exists (Miletic & Gilbert, 2020).

To combat such harms, the moral companion fosters open-ended solutions that broaden the cognitive and emotional space for human moral action. This entails how, in the absence of a clear solution, the companion remains neutral or "uncertain" on the required moral norm. Naturally, this doesn't entail that the companion cannot utilize both its affective and ethical skills to focus the moral deliberation of the human agent on finding possible solutions. It only means that the agent cannot, being symbiotic, make a decision that transcends its symbiotically attuned capacities. As I have previously noted, the machine is not and never will be human. But it can help humans be better and more responsible humans. Here, then, the companion remains „uncertain" (Bogosian, 2017; Kaplan et al. 2018)[102], or „humble" acknowledging that for some moral problems, there simply are no quick or easy solutions.

Conclusively, this entails that the first top layer includes only those ethical frameworks which are without a doubt improving the ethical decision-making for the end-user. It is extremely hard, at this moment in time, to prognosticate specifically what kinds of ethical policies we should implement here. The field of machine ethics is a new field of research, still gaining traction, but as (Gordon, 2020) argues we can adopt the proscription approach – what the artificial agent should never do, or avoid doing, rather than ethics of prescription – what it should do:

[102] "Instead of forcing a decision, it is important for such a system to be able to characterize its uncertainty relative to the similarity of the current observations to the training data, and to explain the uncertainty to a human decision maker... By isolating the uncertainty, the system can then explain the source and nature of the uncertainty to the user. This enables the human to focus his/her reasoning strengths to the special cases that need extra attention." (Kaplan et al., 2018).

> "Machine ethics should 'focus on the similarities between the various ethical approaches in order to gain a better understanding of what one should *not* do. In other words, one should try to spell out in more detail not how moral agents should act but, rather, what they should avoid doing'. This approach has an analogy in the context of moral enhancement, where a similar question—namely, which ethical approach should be used as a default position to make human beings more moral—has been raised. One reasonable answer to this question is to try to agree 'on a binding list of things we want to *avoid* in our moral lives' and to create machines that will act accordingly while, at the same time, permitting them to make moral decisions within a framework that allows for different but equally acceptable moral solutions." (Gordon, 2020, p. 153).

In this regard, as a general rule of thumb, the experts and designers could follow the simple „less is more" principle, or „when in doubt, leave it out". By focusing only on the necessary ethical rules in the first layer, the overall model allows for the plurality of moral choice and fine-tuned augmentation within the second two layers. This leaves us with the second two layers, the parent/caretaker and the user level.

Here I propose to utilize the process of reflective equilibrium.[103] Beneficially, this doesn't only entail a narrow reflective equilibrium but also a wide one, as our companion can rely on the

[103] As (Bengson et al., 2019) defines: "When constructing a theory, a theorist ought to achieve coherence between various particular judgments (e.g., considered judgments regarding specific cases) and beliefs in general principles (e.g., universally quantified propositions) that address all of the central questions about the domain, through a reflective process of modification, addition, and abandonment of either the particular judgments or principles in case of conflict (with each other, or with any of one's other relevant convictions). The best theory is the one that achieves such coherence to the highest degree relative to rivals." (Bengson et al., 2019, p. 412).

practically applicable knowledge coming from the "background theories" of developmental psychology and educational sciences.

Consider these two examples. The parent may tell the companion: "Let John (the child) know that…". To this the companion may respond by reflecting the user's input: "You might consider X". The parent modifies her advice: "Ok, let John know that Y. Or, "John is playing aggressive video games, should I stop him?" To this, the companion responds by providing relevant psychological data on video games and aggressiveness (for instance a video link, or an article).[104]

Naturally, if further developed the companion could not only initiate the reflective equilibrium process, it could also provide more finely-tuned assistance to achieve coherence in the parent's moral judgments. For instance, in the first step, is capable of detecting a strong emotional disturbance in the parent's voice (for instance anger), the companion can propose that the moral judgment at hand ought to be made in a clear and non-biased manner. In the second step of the process, the companion can assist in the formulation of a generally applicable ethical principle by offering examples from its ethical repertoire: "The child's safety should be paramount" or by pointing out some of the previously utilized principles: "Insofar, you have made John always return home by 8:00 PM", or by consulting available examples: "This percentage of parents has limited their children's stay by 8:00 PM". Importantly, here, the companion can also utilize knowledge coming from "background theories" for a possible revision of the ethical principle, for instance: "The statistics of public safety at night is X", or "In his age, it is usual for children to initiate first steps in socialization", or more formally "Consider this information from developmental

[104] Here I am purposefully providing realistic examples which are, in some regard, already being utilized by some of the existing AI companions on the market (for instance Woebot, which I have personally utilized).

psychology". The point of the symbiotic companion is then here to help the parent or caretaker produce better moral judgments, by warning of emotional disturbances and empowering their moral deliberation.

Importantly, the moral companion cannot appreciate inputs from a parent or a caretaker if that input breaks some of its basic constraints.[105] For instance, the AI can be advised to remind the child not to share family secrets around the Sunday meal or the command not to speak with strangers, or enter into a strangers' car. Also, parents could instruct the AI companion to tutor the child on specific moral scenarios but always within the existing fundamental constraints. For instance, the AI could advise the child to be diligent in its education or to avoid harmful content on the internet[106]. Additionally, it can be expected that the input provided by the parents and the user herself will be, at least, partially enculturated. For this reason, the socio-cultural values which do not negate the firm standards above can also be taken into consideration. For example, social robotics already adheres to existing socio-cultural practices and fundamental socio-cultural tenets (for instance, in Japan that of harmonious coexistence 和).

Finally, the level of the user. The ethical model of the user-level aims to establish that the user not only retains her moral autonomy but is also practically augmented in everyday contexts. It does so

[105] For the most recent computational approaches see work of (Loreggia et al., 2018). Here, authors show how the use of CP-nets (ceteris caribus) to model the user's subjective preferences and the ethical principles (moral constraints). If the user's choices (preferences) suggest actions that are breaching the ethical principle, the ethical condition is triggered and the system advises actions that remain within the ethical threshold.

[106] Work of Balakrishnan et al. (2019) is based on reinforcement learning. Here, the agent is restricted, by "a set of *behavioral constraints* that are independent of the reward function. For instance, a parent or guardian group may want a movie recommender system (the agen)t to not recommend certain types of movies to children, even if the recommendation of such movies could lead to a high reward." (Balakrishnan et al., 2019).

by providing normative advice ingrained within the pre-programmed hierarchy of the top, expert, level, and aligned to parental level.[107] As the companion learns of the user's moral preferences, it should aim to build a fine-tuned ethical alignment, within the aforementioned ethical frameworks. As these are purposefully coarse-grained, the plurality of the user's moral choice is not endangered. In this regard, different moral perspectives, that do not negate the basic ethical constraints, can be accepted by the companion. Also, as noted, artificial companions are capable of strongly contributing to the establishment of a narrow and wide moral equilibrium. In this sense, the companion directly improves moral deliberation, as one cannot make better moral judgments if one is not cognizant of the moral beliefs affecting those judgments nor oblivious to other moral possibilities.[108]

I envision that this entails several possible outcomes. Here I point to three such possibilities. At the first level of complexity, the companion could adopt specific moral principles and include this within its ethical model, as it generates advice for the user similarly to prior mentioned examples. Another, more complex, approach entails the utilization of moral characters and not only the implementation of specific moral principles that the companion takes into consideration. With the availability of market opportunities, different moral characters might be offered for download and

[107] This entails that the advice given to the user provides morally normative reasons where the strength of normativity between the upper two levels is not equal. For instance: ”Your mother wants you to remember that…” differs from “I am unable to do that due to my ingrained ethical character, or ethical programming”.

[108] As Rawls points out: „From the standpoint of moral theory, the best account of a person’s sense of justice is not the one which fits his judgments prior to his examining any conception of justice, but rather the one which matches his judgments in reflective equilibrium. As we have seen, this state is one reached after a person has weighed various proposed conceptions and he has either revised his judgments to accord with one of them or held fast to his initial convictions (and the corresponding conception). (Rawls, 1971, p. 43).

utilization by the companion. Naturally, these „moral characters“ can and should be open for evaluation through different kinds of applicable[109] standards. (Winfield, 2019).

The third, possibility entails the AI companion generating a type of moral character from scratch, without any additional input but the foundational ethical principles, the caretakers/parents' advice, and the user's moral inputs. This approach might be the most complex solution not just due to technology but also ethical reasons. Concretely, if the companion learns only from our „own moral parameters the moral criteria with which we would assess its responses are the same as those it has used to provide such responses” (Giubilini & Savulescu, 2018). This entails that the companion finds itself in an effective loop-hole and cannot augment the process of moral deliberation with novel and challenging information.

However, this doesn’t necessarily have to be the case as the promising new research on artificial intuition (Perez, 2018) (based on reinforcement learning and game theory techniques) shows. David Silver, as reported, explains how this approach (exemplified in the case of Alpha Go Zero (Silver, 2017), doesn’t utilize either human-generated data or human expertise for a specific domain” (Knight, 2017) Here, the constraints of human knowledge are removed, and the machine

[109] For instance, a witty example is included in the movie “Looking for Eric”, where the protagonist engages in improvised, but therapeutical, conversation with the persona of Eric Cantona. The point to take here is – humans often do not require much of a “touch” to get moving into a positive direction. But that “touch” has to be personalized. In this sense for many healthy humans the only thing required is for someone to listen to you and respond with a good reflection, advice. Likewise the advice given doesn’t have to be a reference – “do exactly this”. It can also be a vector, - “Do something surprising”. In this sense, symbiotic machines do not have to provide a fixed navigational point for the human to navigate to. Rather, they could simply show the path, which has no precise end-point, but rather leaves that determination of that end-point to the human. In this sense, it fully respects human autonomy and leaves the reigns of living our actions out in our own hands.

finds the best possible solution fully on its own. This was utilized in examples of other complex gaming environments, for instance, the OpenAI Dota2 team (Berner, 2019) or the StarAlpha (Vinvals, 2019). But, complex action in gaming environments is not the only kind of intuitive analysis these systems can achieve as already, in criminal investigations, the "artificial intuition reasoning model is used to analyze a crime investigation case" by constructing its very own inferential process. (Liu et al., 2019). What these breakthroughs showcase is the ability of machines to generate novel and useful ideas that challenge the existing human knowledge in a specific domain. And in doing so, achieve creative (Newell et al., 1962) super-human performance. In this regard, machines withhold the theoretical possibility to create ethical models, on their own, only from basic ethical principles and user-generated input.[110]

However, notwithstanding the generation of ethical models in use, the wide-spread application of such models will have to be regulated, in one way or another. For this reason, as a conclusion of our elaboration on AI companions, I finally engage the question of distribution. Here, most probably, the point of breaking spears for companion augmentation is predominantly neither technical nor ethical but rather political.

5.4. Companion distribution

In this final part, I explore how a wide distribution of AI companions can exacerbate economic and political *power asymmetries* which can effectively threaten the basic tenets of our democratic

[110] As Falon Fatemi, as reported, concludes. " In the future, everyone should have their own artificial intuition agent whose job is to identify opportunities to you and put those in front of you before you even know to search for them. Whether it's "here's the next book you should read that is going to change your life" or "here is the next job opportunity you should take a look at," I think that's longer term where this can go, and really empower all of us." (Newcomb, 2019).

communities (Yeung, p. 7). In this exposition, I will utilize three possible policy approaches that could be adopted as a reaction to the development of companion technology. Inspired by Sarewitz and Karas (2006) I call these *the* affirmative approach, the prohibitory approach, and the regulatory approach. To engage the issue of power asymmetry, and warrant the companion's distribution beneficial outcome I propose the adoption of the regulatory approach.

I start with the affirmative standpoint which accepts that the individual's citizen's freedom to accept the assistance of the AI companion is based on personal judgment established by the right of autonomy within a liberal democracy. For children, this decision would be made by their parents. Here then, the government would not stop the distribution of companions to those that express the need for it. This would leave the dynamics of the free market to regulate the distribution and use of AI companions similarly to how many contemporary social networks remained unregulated in the early periods of their distribution. However, the danger lying here is that powerful interest groups such as Big Tech corporations or ideological groups can find ways to quickly push forward their interests, over the public interest, especially since we have in mind how artificial intelligence is already a multi-billion interest field. And when the public interest, as expressed in legitimate democratic processes, becomes a puppet of interest groups[111] democracy has effectively stopped functioning. Unfortunately, this kind of realistic danger can become even worse with the introduction of personal AI advisors, and the social and political influence Big Tech companies might exert in the process.

[111] «When the prince stops governing the state in accordance with the laws, and usurps the sovereign power then something remarkable happens: the government doesn't contract, but the state does; I mean that the great big state is dissolved, and another state is formed within it, composed solely of the members of the government and relating to the rest of the people as their master and tyrant.» The Social Contract 3,10. (Rousseau, 2018).

This is the problem of power asymmetry, where a wide and unregulated distribution of AI advisors, designed by Big Tech, Big Data firms, can collect tremendous amounts of citizen's data and utilize that data to influence individual behavior. As the Wagner Study points:

> „The increasing use of automation and algorithmic decision-making in all spheres of public and private life is threatening to disrupt the very concept of human rights as protective shields against state interference. The traditional asymmetry of power and information between state structures and human beings is shifting towards an asymmetry of power and information between operators of algorithms (who may be public or private) and those who are acted upon and governed." (Wagner, 2017, p. 33).

Here then we are not only dealing with the danger of having a private group influencing democratically elected officials to serve their own rather than public interests. We are having a parallel system of behavioral influence, direct and intimate, and with the power to "undermine democratic processes, human deliberation and democratic voting systems" (AI HLEG, 2019, p. 11). In the worst-case scenario, this kind of influence over the democratic structure could be done far from public scrutiny, and without the ability for redress. As the recent study, commissioned by the Council of Europe states:"

> „These systems will invariably reflect the values and value priorities of the system and its developers and might not be aligned with the collective values of the public or the democratic and constitutional values that human rights are designed to serve. Yet, even in relation to AI systems that directly affect and interface with the public, citizens and other affected groups and organizations will typically not be given any meaningful opportunity to participate in identifying these values or value trade-offs that these systems are configured

> to reflect...Given their widespread effects, the determination of those values should be subject to democratic participation and deliberation rather than being resolved privately by private providers motivated by commercial self-interest.“ (Yeung, 2018, p. 37).

Unfortunately, the fear of such a development, exacerbated by a lack of design transparency from the AI developers, could easily create a strong backlash against the technology’s otherwise beneficial outcome. This is the prohibitory attitude. Here the governments’ first approach would be to ban companion distribution and limit the possibility of Big Tech firms to conglomerate social and political power by exerting behavioral influence over developing children. Such a stand could be easily fueled by those public groups which take that the current status of human nature, human development, and social relations stands endangered by companion technologies. As such they could argue that a widespread companion distribution changes the meaning and practice of parenting, education, and human development and would motivate the government to prohibit the distribution by adopting an AI-conservative stand. (I.e. “We do not want our children to be raised by robots!”)

To combat the danger of power asymmetry and engage the prohibitory critics, I propose the utilization of the regulatory approach based on the, prior espoused, realistic technological optimism. This third option argues that the utilization of companion technology promises many benefits (including moral augmentation), both for individuals and society as a whole, but these are not guaranteed without active governmentally approved regulation. Potential risk management, protection of human rights and dignity, social balance, are just some of the critical points that are here considered. Again, similarly to the situation we already had in Europe, Big Tech firms might get away with some sudden implementations of AI products, but it can be safely expected that a

wide and firm stand on regulations would soon ensue if necessary ("Press corner", 2019). Different political groups could accept this kind of regulatory stance.

However, to ensure regulatory mechanisms, and the beneficial outcome of companion distribution, the role of non-governmental expert bodies in the decision making process remains essential. As such, additional effort must be invested so that when the time for a political decision comes, the arguments are clear and accessible to the entire public especially when we think how the prospect of AI companions growing up with human children opens up numerous, and prior unprecedented, ideological battle-grounds. However, we have to have in mind that the decision-making process cannot be solely related to the circle of experts. A couple of reasons can be listed to support this conclusion.

First, not all experts reside in the governmentally chosen expert teams. As witnessed in the 2020 COVID pandemics, many noted experts within the public are capable of stimulating and guiding the public discussion and who are, most importantly, not tied or related to political goals or partisanship. Having independent experts freely voicing their opinions can only add valuable dynamics to the discussion. Especially when we have in mind that governmentally approved experts might be the focus of political opponents' ire – justifiably or not – which could cast into doubt their expert evaluations. Thus, the public's non-partisan involvement remains significant for the strengthening of democratic and social cohesion on this matter. Especially, when we have in mind how the public, as the end-user, is here not merely the end receiver of augmentation but also the economic entity that funds scientific and philosophical research. Here then, the public has to remain involved as it is both «the hand that feeds and receives» and even more so – as the engaged political policies directly impact human bearing and children development. The role of

expert groups is then, arguably, to inform and educate the public to allow a healthy and profound public discussion but not to cast the decisive vote on the technology's distribution – similarly to the practice of AI "guidelines". Lastly, including the public in the decision process also prohibits rushing the process of distribution, and does not allow for the crucial decisions to be made in secrecy or haste. This naturally puts the role of the public in the focus as a potent safeguard from the private, socially damaging, goals of interest groups. However, the public may also be manipulated. For instance, if the referendum as a public decision-making process pushes an ad-hoc response initiated by a prohibition group rather than a response to a thorough public discussion. Again, appeals to emotions and political partisanship might be utilized to promote a prohibitory stance, rather than a reasoned and well-informed public discussion. For this reason, it is of the utmost importance to bring the debate on AI companions to the public in a transparent and well-informed way.

Nevertheless the outcome, we have to remain cognizant of power asymmetries, capable of jeopardizing individual rights and democratic processes. As such, manipulations by interest groups are to be monitored and controlled when necessary. It is vital to prioritize active citizenship as an essential safeguarding mechanism, to ensure the beneficial outcome of companion technology. As Francois Chollet concludes:

„We should build AI to serve humans, not to manipulate them for profit or political gain. What if...algorithms didn't operate like casino operators or propagandists? What if instead, they were closer to a mentor or a good librarian, someone who used their keen understanding of your psychology — and that of millions of other similar people — to recommend to you that next book that will most resonate with your objectives and make you grow. A sort of navigation tool for your

life — an AI capable of guiding you through the optimal path in experience space to get where you want to go. Can you imagine looking at your own life through the lens of a system that has seen millions of lives unfold? Or writing a book together with a system that has read every book? Or conducting research in collaboration with a system that sees the full scope of current human knowledge?“ (Chollet, 2019)

CONCLUSION

> "Then we are living in a place abandoned by God," I said, disheartened. "Have you found any places where God would have felt at home?" William asked me, looking down from his great height." — Umberto Eco, The Name of the Rose

When Benedict of Nursia, one of the most important cultural and moral figures of Europe's Middle Ages, set out to create his "Rules of Saint Benedict" and form the Benedictine order he did not want to start something grand. Rather, he perceived his endeavor as a modest one, one which did not ask too much from the individual human and one which, in all actuality, did not strain the social structure by its implementation. As such, he created a small step, but one which resulted in lasting beneficial outcomes.

I first heard of this narrative, while visiting the Kornelimünster Abbey in Aachen (Germany) in the Summer of 2007. Close to that time, I have also come into my first contacts with transhumanism, through the works of Max More and Nick Bostrom, and I felt deeply engrossed by their vision. The prospect of improving human nature, our limitations, and frailties, through the power of technology, has deeply resonated with me ever since. I found it highly motivating that after thousands of years of tries and failures, human civilization has come to the point where it is becoming capable to improve its nature, in a defined and observable way. It seemed to me that the human civilization could finally resolve its internal moral maladies, its weak will, the cognitive inconsistencies, the plaguing emotional disturbances. Yet, Benedict's idea of starting small and modest was a thought I seemed to never really get rid of. It remained somewhere in the back of mind, as a powerful moral intuition, some sort of a small, feint, the voice of wisdom which seemed to say: "Could it be possible to utilize technology in a modest, safe, way but to still achieve great

and worthy moral improvement outcomes?" Today, I believe the answer to this question is affirmative, as I have tried to elaborate in this work.

Here I have provided reasons for the proposal of symbiotic moral augmentation, a proposal that builds upon the existing moral research achievements and provides an advancement over the available approaches of AI-related means of moral enhancement. I have argued that the symbiotic moral augmentation is capable of achieving a fine-tuned and ethical-moral improvement that is aligned with the child's development and moral autonomy and is safely applicable within contemporary democratic standards. I have also anchored the symbiotic means on realistic technological grounds, explicated the ethical nature of artificial companions, elaborated on the sensitivity towards the child's developmental phases, and illustrated how the paradigm of symbiotic companions improves acceptance. In all of this, I remained steady to point out how the means to achieve this vision, albeit modestly, is already in our hands. AI systems, especially AI assistants, are already assisting human decisions, affecting our moods and emotions, and as time inevitably moves further forward it is safe to presume that its prowess will rise and our relation with AI companions will grow more intimate. However, as I have tried to constantly warn of - the decision on what we want to do with AI technology, the ethical path where we want to take our AI systems to is ours to make. Here, I have aimed to corroborate how, in the cacophony of AI-related ethical and political issues, the symbiotic design offers a clear perspective on a socially implementable approach towards the development of empowering AI technologies which foster human flourishing.

Again, such a vision may start modestly, but it envisions a grand goal. As for the first time in human history, humanity is becoming capable to achieve the prior unimaginable – the creation of artificial

companions. It seems only reasonable to conclude that the vision of symbiotic, morally augmentative, AI companions, is worthy to be pursued. Hopefully, this work will establish the initial justification of symbiotic moral companions as means of moral augmentation and will inspire the right, ethical, direction of moral AI development. To this kind of beneficial future, I dedicate my proposal.

REFERENCES

1. Abel, D., MacGlashan, J., & Littman, M. L. (2016). Reinforcement Learning as a Framework for Ethical Decision Making. In *AAAI Workshop: AI, Ethics, and Society* (Vol. 92).

2. Abney, K. (2012). Robotics, ethical theory, and metaethics: A guide for the perplexed. *Robot ethics: The ethical and social implications of robotics*, 35-52.

3. Agar, N. (2008). *Liberal eugenics: In defense of human enhancement*. John Wiley & Sons.

4. Amodei, D., & Clark, J. (2016). Faulty reward functions in the wild. *URL: https://blog. openai.com/faulty-reward-functions.*

5. Anderson, J. (2003). Autonomy and the authority of personal commitments: From internal coherence to social normativity. *Philosophical Explorations*, 6(2), 90-108.

6. Anderson, J., & Rainie, L. (2020). *Many Tech Experts Say Digital Disruption Will Hurt Democracy*. Pew Research Center.

7. Anderson, M., & Anderson, S. L. (2007). Machine ethics: Creating an ethical intelligent agent. *Ai Magazine*, 28(4), 15-15.

8. Anderson, M., & Anderson, S. L. (Eds.). (2011). *Machine ethics*. Cambridge University Press.

9. Anderson, M., Anderson, S. L., & Armen, C. (2006, August). MedEthEx: a prototype medical ethics advisor. In *Proceedings of the national conference on artificial intelligence* (Vol. 21, No. 2, p. 1759). Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

10. Anderson, S. L., & Anderson, M. (2011, August). A prima facie duty approach to machine ethics and its application to elder care. In *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*.

11. Arkin, R. (2009). *Governing lethal behavior in autonomous robots*. CRC Press.

12. Arkin, R. C. (2018). Ethics of Robotic Deception [Opinion]. *IEEE Technology and Society Magazine*, 37(3), 18-19.

13. Armstrong, S. (2015, April). Motivated value selection for artificial agents. In *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*.

14. Asaro, P. M. (2006). What should we want from a robot ethic?. *The International Review of Information Ethics*, 6, 9-16.

15. Bainbridge, L. (1983). Ironies of automation. In *Analysis, design and evaluation of man–machine systems* (pp. 129-135). Pergamon.

16. Balakrishnan, A., Bouneffouf, D., Mattei, N., & Rossi, F. (2019, July). Incorporating behavioral constraints in online AI systems. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 33, pp. 3-11).

17. Barrett, L. F. (2017). *How emotions are made: The secret life of the brain*. Houghton Mifflin Harcourt.

18. BD Editors. (2017, April 29). Parasitism. *Biology Dictionary*. https://biologydictionary.net/parasitism/

19. BD Editors. (2019, April 8). Symbiosis. *Biology Dictionary*. https://biologydictionary.net/symbiosis/

20. Bello, P., & Bringsjord, S. (2013). On how to build a moral machine. *Topoi*, 32(2), 251-266.

21. Berner, C., Brockman, G., Chan, B., Cheung, V., Dębiak, P., Dennison, C., ... & Józefowicz, R. (2019). Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680.*

22. Berreby, F., Bourgne, G., & Ganascia, J. G. (2015, November). Modelling moral reasoning and ethical responsibility with logic programming. In *Logic for programming, artificial intelligence, and reasoning* (pp. 532-548). Springer, Berlin, Heidelberg.

23. Berridge, K. C. (2018). Evolving concepts of emotion and motivation. *Frontiers in Psychology*, 9, 1647.

24. Bogosian, K. (2017). Implementation of moral uncertainty in intelligent machines. *Minds and Machines*, 27(4), 591-608.

25. Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in neural information processing systems* (pp. 4349-4357).

26. Bonnefon, J. F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science*, 352(6293), 1573-1576.

27. Borenstein, J., & Arkin, R. (2019). Robots, ethics, and intimacy: The need for scientific research. In *On the cognitive, ethical, and scientific dimensions of artificial intelligence* (pp. 299-309). Springer, Cham.

28. Bostrom, N. (2005). Transhumanist values. *Journal of philosophical research*, 30(Supplement), 3-14.

29. Bostrom, N. (2008). Letter from utopia. *Studies in Ethics, Law, and Technology*, 2(1).

30. Bostrom, N. (2016). *Superintelligence: Paths, Dangers, Strategies*, Reprint ed.

31. Bostrom, N., More, M., Vita-More, N., Pierce, D., & Morrow, T. (2009). Transhumanist declaration. *Humanity+*. http://humanityplus. org/philosophy/transhumanist-declaration.

32. Brey, P. (2009). Human enhancement and personal identity. In *New waves in philosophy of technology* (pp. 169-185). Palgrave Macmillan, London.

33. Bringsjord, S., & Taylor, J. (2012). The divine-command approach to robot ethics. *Robot ethics: The ethical and social implications of robotics*, 85-108.

34. Bringsjord, S., Ghosh, R., & Payne-Joyce, J. (2016). Deontic counteridenticals. *Agents (EDIA),* 2016, 40-45.

35. Brockman, J. (2014). The Myth Of AI | Edge.org. https://www.edge.org/conversation/the-myth-of-ai#26015

36. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Agarwal, S. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

37. Buchanan, A., & Powell, R. (2018). *The evolution of moral progress: a biocultural theory*. Oxford University Press.

38. Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1), 2053951715622512.

39. Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334), 183-186.

40. Campolo, A., Sanfilippo, M., Whittaker, M., & Crawford, K. (2017). AI now 2017 report. *AI Now Institute at New York University.*

41. Cave, S., Nyrup, R., Vold, K., & Weller, A. (2018). Motivations and risks of machine ethics. *Proceedings of the IEEE*, 107(3), 562-574.

42. Challen, R., Denny, J., Pitt, M., Gompels, L., Edwards, T., & Tsaneva-Atanasova, K. (2019). Artificial intelligence, bias and clinical safety. *BMJ Quality & Safety*, 28(3), 231-237.

43. Chen, D. L. (2019). Judicial analytics and the great transformation of American Law. *Artificial Intelligence and Law*, 27(1), 15-42.

44. Chollet, F. (2018). What worries me about AI - François Chollet. *Medium*. https://medium.com/@francois.chollet/what-worries-me-about-ai-ed9df072b704

45. Christen, M., & Narvaez, D. (2012). Moral development in early childhood is key for moral enhancement. *AJOB Neuroscience*, 3(4), 25-26.

46. Christman, J., & Zalta, E. N. (2015). Autonomy in Moral and Political Philosophy. *The Stanford encyclopedia of philosophy.*

47. Chung, M., Fortunato, G., & Radacsi, N. (2019). Wearable flexible sweat sensors for healthcare monitoring: a review. *Journal of the Royal Society Interface*, 16(159), 20190217.

48. Ciolacu, M., Tehrani, A. F., Beer, R., & Popp, H. (2017, October). Education 4.0—Fostering student's performance with machine learning methods. In *2017 IEEE 23rd International Symposium for Design and Technology in Electronic Packaging (SIITME)* (pp. 438-443). IEEE.

49. Code, L. (1984). Toward a Responsibilist' Epistemology. *Philosophy and phenomenological research*, 45(1), 29-50.

50. Coleman, K. G. (2001). Android arete: Toward a virtue ethic for computational agents. *Ethics and Information Technology*, 3(4), 247-265.

51. Coplan, A. (2011). Will the real empathy please stand up? A case for a narrow conceptualization. *The Southern Journal of Philosophy*, 49, 40-65.

52. Crowell, S. (2020.) Existentialism. *The Stanford Encyclopedia of Philosophy*.

53. Cutuli, D. (2014). Cognitive reappraisal and expressive suppression strategies role in the emotion regulation: an overview on their modulatory effects and neural correlates. Frontiers in systems neuroscience, 8, 175.

54. Danaher, J. (2020). Robot Betrayal: a guide to the ethics of robotic deception. *Ethics and Information Technology*, 1-12.

55. Das, A. K., Ashrafi, A., & Ahmmad, M. (2019, February). Joint Cognition of Both Human and Machine for Predicting Criminal Punishment in Judicial System. In *2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS)* (pp. 36-40). IEEE.

56. Deng, B. (2015). Machine ethics: The robot's dilemma. *Nature News*, 523(7558), 24.

57. Dennis, L. A., Fisher, M., & Winfield, A. F. (2015). Towards verifiably ethical robot behaviour. *arXiv preprint arXiv:1504.03592.*

58. Dennis, L., Fisher, M., Slavkovik, M., & Webster, M. (2016). Formal verification of ethical choices in autonomous systems. *Robotics and Autonomous Systems*, 77, 1–14.

59. Díaz Boladeras, M. (2017). Bonding with robotic pets. Children's cognitions, emotions and behaviors towards pet-robots. *Applications in a robot assisted quality of life intervention in a pediatric hospital*.

60. Dietterich, T. G., & Horvitz, E. J. (2015). Rise of concerns about AI: reflections and directions. *Communications of the ACM*, 58(10), 38-40.

61. Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1), 114.

62. Dignum, V. (2018). Ethics in artificial intelligence: introduction to the special issue. *Ethics and Information Technology*, 20(1), 1–3.

63. Dormandy, K. (2018). Epistemic Authority: Preemption or Proper Basing?. Erkenntnis, 83(4), 773-791.

64. Douglas, T. (2013). Moral enhancement via direct emotion modulation: a reply to John Harris. *Bioethics*, 27(3), 160-168.

65. Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1), eaao5580.

66. Drew, L. (2019). The ethics of brain-computer interfaces. *Nature*, 571(7766), S19-S19.

67. Drubin, C. (2019). Building Trusted Human-Machine Partnerships. *Darpa.Mil*. https://www.darpa.mil/news-events/2019-01-31

68. Drweesh, Z. T., & Al-Bakry, A. (2019). Medical Diagnosis Advisor System: A Survey. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, 8(1).

69. Dubljević, V. (2013). Cognitive enhancement, rational choice and justification. *Neuroethics*, 6(1), 179-187.

70. Dubljević, V., & Racine, E. (2017). Moral enhancement meets normative and empirical reality: assessing the practical feasibility of moral enhancement neurotechnologies. *Bioethics*, 31(5), 338-348.

71. Edirisinghe, C., Cheok, A. D., & Khougali, N. (2017, December). Perceptions and responsiveness to intimacy with robots; a user evaluation. In *International conference on love and sex with robots* (pp. 138-157). Springer, Cham.

72. Edwards, A. (2019.). Times Spock and McCoy Were Basically an Old Married Couple. Ranker. Retrieved 2020, from https://www.ranker.com/list/mccoy-spock-star-trek-moments/aaron-edwards

73. Elsayed, G., Shankar, S., Cheung, B., Papernot, N., Kurakin, A., Goodfellow, I., & Sohl-Dickstein, J. (2018). Adversarial examples that fool both computer vision and time-limited humans. In *Advances in Neural Information Processing Systems* (pp. 3910-3920).

74. Engelbart, D. C. (1962). *Augmenting human intellect: A conceptual framework*. Menlo Park, CA.

75. Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C., & Venkatasubramanian, S. (2018, January). Runaway feedback loops in predictive policing. In *Conference on Fairness, Accountability and Transparency* (pp. 160-171).

76. Etzioni, A., & Etzioni, O. (2017). Incorporating ethics into artificial intelligence. *The Journal of Ethics*, 21(4), 403-418.

77. Fantl, J., (2019). Knowledge How. *The Stanford Encyclopedia of Philosophy*.

78. Feinberg, J. (1989). *Autonomy/The Inner Citadel–Essays on Individual Autonomy*, ed. John Christman.

79. Ferguson, A. G. (2016). Policing predictive policing. *Wash. UL Rev.*, 94, 1109.

80. Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Schafer, B. (2018). AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Minds and Machines,* 28(4), 689-707.

81. Ford, M. (2015). *Rise of the Robots: Technology and the Threat of a Jobless Future*. Basic Books.

82. Fukuyama, F. (2003). *Our posthuman future: Consequences of the biotechnology revolution*. Farrar, Straus and Giroux.

83. Gao, W., Emaminejad, S., Nyein, H. Y. Y., Challa, S., Chen, K., Peck, A., ... & Lien, D. H. (2016). Fully integrated wearable sensor arrays for multiplexed in situ perspiration analysis. *Nature*, 529(7587), 509-514.

84. Gatys, L. A., Ecker, A. S., & Bethge, M. (2015). A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*.

85. Gips, J. (1994). Toward the ethical robot. *Android Epistemology*. MIT Press. pp. 243—252.

86. Giubilini, A., & Savulescu, J. (2018). The artificial moral advisor. The "ideal observer" meets artificial intelligence. *Philosophy & technology*, 31(2), 169-188.

87. Goodall, N. J. (2014). Ethical Decision Making during Automated Vehicle Crashes. *Transportation Research Record: Journal of the Transportation Research Board*, 2424(1), 58–65.

88. Gordon, J. S. (2020). Building moral robots: ethical pitfalls and challenges. *Science and engineering ethics*, 26(1), 141-157.

89. Grcic, J. (2007). Hobbes and Rawls on Political Power, *Etica & Politica / Ethics & Politics* IX, 371-392.

90. Grcic, J. M. (1985). Rawls and rousseau on the social contract. *Auslegung: a journal of philosophy*, 12(1), 70-81.

91. Greco, J., & De Sa, L. P. (2018). *Epistemic Value*. Oxford: Oxford University Press, 2009).

92. Green, B., & Hu, L. (2018, July). The myth in the methodology: Towards a recontextualization of fairness in machine learning. In *Proceedings of the machine learning: the debates workshop*.

93. Greene, J., Rossi, F., Tasioulas, J., Venable, K. B., & Williams, B. C. (2016, February). Embedding Ethical Principles in Collective Decision Support Systems. In *Aaai* (Vol. 16, pp. 4147-4151).

94. Grgic-Hlaca, N., Zafar, M. B., Gummadi, K. P., & Weller, A. (2018, February). Beyond Distributive Fairness in Algorithmic Decision Making: Feature Selection for Procedurally Fair Learning. In AAAI (pp. 51-60).

95. Groll, D. (2012). Paternalism, respect, and the will. *Ethics*, 122(4), 692-720.

96. Guarini, M. (2013). Introduction: Machine Ethics and the Ethics of Building Intelligent Machines. *Topoi*, 32(2), 213–215.

97. Hadhazy, A. (2017). Biased Bots: Artificial-Intelligence Systems Echo Human Prejudices. Princeton University, April, 18.

98. Hagendorff, T. (2020). The ethics of Ai ethics: An evaluation of guidelines. *Minds and Machines*, 1-22.

99. Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in neural information processing systems* (pp. 3315-3323).

100. Haselton, T. C. F. (2018, June 6). Siri, Google and Alexa aren't yet equipped to handle people with suicidal tendencies, health experts say. *CNBC*. https://www.cnbc.com/2018/06/06/siri-alexa-google-assistant-responses-to-suicidal-tendencies.html

101. Hauskeller, M. (2012). My brain, my mind, and I: some philosophical assumptions of mind-uploading. *International journal of machine consciousness*, 4(01), 187-200.

102. Heersmink, R. (2017). Distributed cognition and distributed morality: Agency, artifacts and systems. *Science and Engineering Ethics*, 23(2), 431-448.

103. Helbing, F. B. D. S. (2017, February 25). Will Democracy Survive Big Data and Artificial Intelligence? *Scientific American*. https://www.scientificamerican.com/article/will-democracy-survive-big-data-and-artificial-intelligence/

104. Helion, C., & Ochsner, K. N. (2018). The role of emotion regulation in moral judgment. *Neuroethics*, 11(3), 297-308.

105. Herbert, C. (2020). An Experimental-Psychological Approach for the Development of Character Computing. In *Character Computing* (pp. 17-38). Springer, Cham.

106. Hernández-Orallo, J., & Vold, K. (2019, January). Ai extenders: The ethical and societal implications of humans cognitively extended by Ai. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 507-513).

107. Hooker, J. N., & Kim, T. W. N. (2018, December). Toward non-intuition-based machine and artificial intelligence ethics: A deontological approach based on modal logic. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 130-136).

108. Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H., & Aerts, H. J. (2018). Artificial intelligence in radiology. Nature Reviews Cancer, 18(8), 500-510.

109. House Of Lords Select Committee. (2018). *Ai in the uk: ready, willing and able*. House of Lords, 36.

110. Howard, D., & Muntean, I. (2017). Artificial moral cognition: moral functionalism and autonomous moral agency. In *Philosophy and computing* (pp. 121-159). Springer, Cham.

111. Jack, R. E., Garrod, O. G., Yu, H., Caldara, R., & Schyns, P. G. (2012). Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences*, 109(19), 7241-7244.

112. Jennings, N. R., Moreau, L., Nicholson, D., Ramchurn, S., Roberts, S., Rodden, T., & Rogers, A. (2014). Human-agent collectives. *Communications of the ACM*, 57(12), 80-88.

113. Johansson, M., Kleinke, S., & Lehti, L. (2017). The digital agora of social media: Introduction.

114. Johnson, R., & Cureton, A. (2019). Kant's moral philosophy. *The Stanford Encyclopedia of Philosophy*.

115. Jordan, N. (1963). Allocation of functions between man and machines in automated systems. *Journal of applied psychology*, 47(3), 161.

116. K9 Burial Protocol | NATIONAL SHERIFFS' ASSOCIATION. Sheriffs.org. (2020). Retrieved 28 August 2020, from https://www.sheriffs.org/programs/articles-white-papers.

117. Kahane, G., & Savulescu, J. (2015). Normal human variation: refocussing the enhancement debate. *Bioethics*, 29(2), 133-143.

118. Kahneman, D., Rosenfield, A. M., Gandhi, L., & Blaser, T. (2016). Noise: How to Overcome the High, Hidden Cost of Inconsistent Decision Making. https.hbr. org/2016/10/noise.

119. Kantowitz, B. H., & Sorkin, R. D. (1987). Allocation of functions. In G. Salvendy (Ed.), *Handbook of human factors* (p. 355–369).

120. Kaplan, L., Cerutti, F., Sensoy, M., Preece, A., & Sullivan, P. (2018). Uncertainty aware AI ML: why and how. *arXiv preprint arXiv:1809.07882*.

121. Karnofsky, H. (2016). Potential Risks from Advanced Artificial Intelligence: The Philanthropic Opportunity. *Open Philanthropy*. Retrieved 28 August 2020, from https://www.openphilanthropy.org/blog/potential-risks-advanced-artificial-intelligence-philanthropic-opportunity.

122. Kim, J., Park, S., & Robert, L. (2019). Conversational Agents for Health and Wellbeing: Review and Future Agendas. Presented at the *Identifying Challenges and Opportunities in Human–AI Collaboration in Healthcare* at the *22th ACM Conference on Computer Supported Cooperative Work and Social Computing* (CSCW 2019), Austin, Texas.

123. Kim, N. G., & Son, H. (2015). How facial expressions of emotion affect distance perception. *Frontiers in psychology*, 6, 1825.

124. Kingwell, K. (2013). Implantable device advises patients with epilepsy of seizure likelihood. *Nature Reviews Neurology*, 9(6), 297-297.

125. Kiverstein, J., & Farina, M. (2011). Embraining culture: leaky minds and spongy brains. *Teorema: Revista Internacional de Filosofía*, 35-53.

126. Kleeman, S. (2016). Here Are the Microsoft Twitter Bot's Craziest Racist Rants. *Gizmodo*. https://gizmodo.com/here-are-the-microsoft-twitter-bot-s-craziest-racist-ra-1766820160. Retrieved October, 22, 2018.

127. Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human decisions and machine predictions. *The quarterly journal of economics*, 133(1), 237-293.

128. Klincewicz, M. (2016). Artificial intelligence as a means to moral enhancement. *Studies in Logic, Grammar and Rhetoric*, 48(1), 171-187.

129. Knight, W. (2017). AlphaGo Zero Shows Machines Can Become Superhuman Without Any Help. *MIT Technology Review*. Retrieved 28 August 2020, from https://www.technologyreview.com/2017/10/18/148511/alphago-zero-shows-machines-can-become-superhuman-without-any-help/.

130. Koay, K. L., Syrdal, D. S., Dautenhahn, K., & Walters, M. L. (2020). A narrative approach to human-robot interaction prototyping for companion robots. *Paladyn, Journal of Behavioral Robotics*, 11(1), 66-85.

131. Leben, D. (2018). *Ethics for robots: How to design a moral algorithm*. Routledge.

132. Lex Fridman. (2019, June 17). Rosalind Picard: Affective Computing, Emotion, Privacy, and Health | Lex Fridman Podcast #24. YouTube. https://www.youtube.com/watch?v=kq0VO1FqE6I

133. Li, Y. (2017). Deep reinforcement learning: An overview. *arXiv preprint arXiv:1701.07274.*

134. Licklider, J. C. (1960). Man-computer symbiosis. *IRE transactions on human factors in electronics*, (1), 4-11.

135. Lilley, S. (2013). Introduction to the Transhumanity Debate. In *Transhumanism and Society* (pp. 1-12). Springer, Dordrecht.

136. Lin, P., Abney, K., & Bekey, G. A. (Eds.). (2012). *Robot ethics: the ethical and social implications of robotics*. Intelligent Robotics and Autonomous Agents series.

137. Liu, S., & He, P. (2019). Artificial Intuition Reasoning System (AIRS) and Application in Criminal Investigations. In *Journal of Physics: Conference Series* (Vol. 1302, No. 3, p. 032032). IOP Publishing.

138. Löffler, D., Schmidt, N., & Tscharn, R. (2018). Multimodal expression of artificial emotion in social robots using color, motion and sound. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction (*pp. 334-343).

139. Lombrozo, T. (2009). Explanation and categorization: How "why?" informs "what?". *Cognition*, 110(2), 248-253.

140. Loreggia, A., Mattei, N., Rossi, F., & Venable, K. B. (2018, December). Preferences and ethical principles in decision making. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 222-222).

141. Lorenz-Spreen, P., Mønsted, B. M., Hövel, P., & Lehmann, S. (2019). Accelerating dynamics of collective attention. *Nature communications*, 10(1), 1-9.

142. Machine Ethics. (2018, July 23). Luciano Floridi - Professor of Philosophy & Ethics of Information, University of Oxford. YouTube. https://www.youtube.com/watch?v=YsLlWvyMysA

143. Malle, B. F., Scheutz, M., & Austerweil, J. L. (2017). Networks of social and moral norms in human and robot agents. In *A world with robots* (pp. 3-17). Springer, Cham.

144. Mancini, G., Biolcati, R., Agnoli, S., Andrei, F., & Trombini, E. (2018). Recognition of facial emotional expressions among italian pre-adolescents, and their affective reactions. *Frontiers in psychology*, 9, 1303.

145. Markoff, J. (2016). *Machines of loving grace: The quest for common ground between humans and robots*. HarperCollins Publishers.

146. McCarthy, J., & Feigenbaum, E. A. (1990). In memoriam: Arthur samuel: Pioneer in machine learning. *AI Magazine*, 11(3), 10-10.

147. Meaney, M. J. (2001). Maternal care, gene expression, and the transmission of individual differences in stress reactivity across generations. *Annual review of neuroscience*, 24(1), 1161-1192.

148. Mehlman, M. J. (2003). Wondergenes: Genetic enhancement and the future of society. Indiana University Press.

149. Mehlman, M. J., & Botkin, J. R. (1998). Access to the genome: the challenge to equality. Georgetown University Press.

150. Menary, R. (2010). Cognitive integration and the extended mind. *The extended mind*, 227-243.

151. Mental Health Chatbot. (2020, August 17). *Woebot*. https://woebothealth.com/

152. Mercer, C., & Trothen, T. J. (Eds.). (2014). *Religion and Transhumanism: The Unknown Future of Human Enhancement: The Unknown Future of Human Enhancement*. ABC-CLIO.

153. Miletić, T. (2020). Military Medical Enhancement and Autonomous AI Systems: Requirements, Implications, Concerns. In Ethics of Medical Innovation, Experimentation, and Enhancement in Military and Humanitarian Contexts (pp. 175-194). Springer, Cham.

154. Miletić, T. & Gilbert, F. (2020). Does Ai Brain Implant Compromise Agency? Examining Potential Harms Of Brain-Computer Interfaces. In The Age of Artificial Intelligence: An Exploration (pp. 253-272). Vernon Press.

155. Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267, 1-38.

156. Miner, A. S., Milstein, A., Schueller, S., Hegde, R., Mangurian, C., & Linos, E. (2016). Smartphone-based conversational agents and responses to questions about mental health, interpersonal violence, and physical health. *JAMA internal medicine*, 176(5), 619-625.

157. Mitchell, W. J., Ho, C. C., Patel, H., & MacDorman, K. F. (2011). Does social desirability bias favor humans? Explicit–implicit evaluations of synthesized speech support a new HCI model of impression management. *Computers in Human Behavior*, 27(1), 402-412.

158. Mittelstadt, B., Russell, C., & Wachter, S. (2019, January). Explaining explanations in AI. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 279-288).

159. Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., & Riedmiller, M. (2013). Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602.*

160. Moor, J. H. (2006). The nature, importance, and difficulty of machine ethics. *IEEE intelligent systems*, 21(4), 18-21.

161. More, M. (2013). The philosophy of transhumanism. The transhumanist reader, 8.

162. Murray, G. (2017, May). Stoic ethics for artificial agents. In *Canadian Conference on Artificial Intelligence* (pp. 373-384). Springer, Cham.

163. Mulligan, D. K., Kroll, J. A., Kohli, N., & Wong, R. Y. (2019). This Thing Called Fairness: Disciplinary Confusion Realizing a Value in Technology. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), 1-36.

164. Nagasawa, M., Mitsui, S., En, S., Ohtani, N., Ohta, M., Sakuma, Y., ... & Kikusui, T. (2015). Oxytocin-gaze positive loop and the coevolution of human-dog bonds. *Science*, 348(6232), 333-336.

165. Narain, J., Quach, T., Davey, M., Park, H. W., Breazeal, C., & Picard, R. (2020, April). Promoting Wellbeing with Sunny, a Chatbot that Facilitates Positive Messages within Social Groups. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1-8).

166. Narvaez, D. (2014). *Neurobiology and the Development of Human Morality: Evolution, Culture, and Wisdom (Norton Series on Interpersonal Neurobiology).* WW Norton & Company.

167. Nass, C., Moon, Y., & Green, N. (1997). Are machines gender neutral? Gender-stereotypic responses to computers with voices. *Journal of applied social psychology*, 27(10), 864-876.

168. Neely, E. L. (2019). The ethics of choice in single-player video games. In *On the Cognitive, Ethical, and Scientific Dimensions of Artificial Intelligence* (pp. 341-355). Springer, Cham.

169. Neto, B. D. S., Silva, V. T., & Lucena, C. J. P. (2011). Nbdi: an architecture for goal-oriented normative agents. In *ICAART 2011*.

170. New York Times Service. (2017, September 29). Self-driving cars, humans must learn a common language. *Boston.Com*. https://www.boston.com/cars/car-news/2017/09/29/self-driving-cars-humans-must-learn-a-common-language

171. Newcomb, A. (2019). Artificial Intuition Wants to Guide Business Decisions. Can It Improve on 'Going With Your Gut'?. *Fortune*. Retrieved 28 August 2020, from https://fortune.com/2019/07/11/artificial-intuition-node-ai-business-decisions/.

172. Newell, A., Shaw, J. C., & Simon, H. A. (1962). The processes of creative thinking. In *Contemporary Approaches to Creative Thinking*, 1958, University of Colorado, CO, US; Atherton Press.

173. Ng, A. (2017, January). Artificial intelligence is the new electricity. In *presentation at the Stanford MSx Future Forum* 2017.

174. Noothigattu, R., Gaikwad, S. N. S., Awad, E., Dsouza, S., Rahwan, I., Ravikumar, P., & Procaccia, A. D. (2017). A voting-based system for ethical decision making. *arXiv preprint arXiv:1709.06692.*

175. Norman, D. A. (2015). The human side of automation. In *Road vehicle automation* 2 (pp. 73-79). Springer, Cham.

176. Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future—big data, machine learning, and clinical medicine. *The New England journal of medicine*, 375(13), 1216.

177. Olsen, K. (2013). English police force sets up retirement plan for dogs. *Pensions & Investments*. 41 (24): 8

178. Pagallo, U. (2017). When morals ain't enough: Robots, ethics, and the rules of the law. *Minds and machines*, 27(4), 625-638.

179. Palermos, O. (2019). *Sorestispalermos.info*. Retrieved 29 August 2020, from https://www.sorestispalermos.info/.

180. Paulo, N., & Bublitz, J. C. (2019). How (not) to argue for moral enhancement: Reflections on a decade of debate. *Topoi*, 38(1), 95-109.

181. Pereira, L. M., & Saptawijaya, A. (2009). Modeling morality with prospective logic. *International Journal of Reasoning-Based Intelligent Systems*, 1(3/4), 209.

182. Pereira, L. M., & Saptawijaya, A. (2016). *Programming machine ethics* (Vol. 26). Cham: Springer.

183. Perez, C. E. (2018). *Artificial Intuition: The Improbable Deep Learning Revolution*. Carlos E. Perez.

184. Persson, I., & Savulescu, J. (2008). The perils of cognitive enhancement and the urgent imperative to enhance the moral character of humanity. *Journal of applied philosophy*, 25(3), 162-177.

185. Persson, I., & Savulescu, J. (2010, December). Moral transhumanism. In *The Journal of Medicine and Philosophy: A Forum for Bioethics and Philosophy of Medicine* (Vol. 35, No. 6, pp. 656-669). Oxford University Press.

186. Persson, I., & Savulescu, J. (2012). *Unfit for the future: the need for moral enhancement*. OUP Oxford.

187. Picard, R. W. (1997). Affective computing MIT press. Cambridge, Massachsusetts.

188. Pinker, S. (2011). *The better angels of our nature: The decline of violence in history and its causes*. Penguin uk.

189. Powers, T. M. (2006). Prospects for a Kantian Machine. *IEEE Intelligent Systems,* 21(4), 46–51. https://doi.org/10.1109/mis.2006.77

190. Preece, A., Harborne, D., Braines, D., Tomsett, R., & Chakraborty, S. (2018). Stakeholders in explainable AI. *arXiv preprint arXiv:1810.00184.*

191. Press corner. European Commission - European Commission. (2019). Retrieved 29 August 2020, from https://ec.europa.eu/commission/presscorner/detail/en/IP_19_1770.

192. Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J. F., Breazeal, C., ... & Jennings, N. R. (2019). Machine behaviour. *Nature, 568(7753)*, 477-486.

193. Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., ... & Lungren, M. P. (2017). Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*.

194. Ramsay, D. S., & Woods, S. C. (2014). Clarifying the roles of homeostasis and allostasis in physiological regulation. *Psychological review*, 121(2), 225.

195. Raus, K., Focquaert, F., Schermer, M., Specker, J., & Sterckx, S. (2014). On defining moral enhancement: a clarificatory taxonomy. *Neuroethics*, 7(3), 263-273.

196. Rawls, J. (1971). *A theory of justice*. Harvard university press.

197. Rawls, J. (1999). *The law of peoples: with, the idea of public reason revisited*. Harvard University Press.

198. Reidsma, D., Charisi, V., Davison, D., Wijnen, F., van der Meij, J., Evers, V., ... & Mazzei, D. (2016, July). The EASEL project: Towards educational human-robot symbiotic interaction. In *Conference on Biomimetic and Biohybrid Systems* (pp. 297-306). Springer, Cham.

199. *Robot ethics: The ethical and social implications of robotics*, 85-108.

200. ROQ 'N ROLLA Music. (2016, March 4). Jebroer - Me Gabber (prod. by Rät N FrikK). YouTube. https://www.youtube.com/watch?v=BgA3zP5PmEY

201. Rosenthal, S., Biswas, J., & Veloso, M. M. (2010, May). An effective personal mobile robot agent through symbiotic human-robot interaction. In *AAMAS* (Vol. 10, pp. 915-922).

202. Rossi, F., & Mattei, N. (2019, July). Building ethically bounded AI. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 33, pp. 9785-9789).

203. Rousseau, J. J. (2018). *Rousseau: The Social Contract and other later political writings*. Cambridge University Press.

204. Russell, S. (2017). *Provably Beneficial Artificial Intelligence* [Ebook]. Retrieved 27 August 2020, from https://people.eecs.berkeley.edu/~russell/papers/russell-bbvabook17-pbai.pdf.

205. Russell, S., & Norvig, P. (2002). *Artificial intelligence: a modern approach*.

206. Sandberg, A. (2013). Morphological Freedom-Why we not just want it, but need it. *The transhumanist reader*, 56-64.

207. Sarewitz, D., & Karas, T. H. (2012). Policy Implications of Technologies for Cognitive Enhancement. *Neurotechnology: Premises, potential, and problems*, 267.

208. Sato, W., Hyniewska, S., Minemoto, K., & Yoshikawa, S. (2019). Facial expressions of basic emotions in Japanese laypeople. *Frontiers in psychology*, 10, 259.

209. Savulescu, J., & Maslen, H. (2015). Moral enhancement and artificial intelligence: Moral AI?. In *Beyond Artificial Intelligence* (pp. 79-95). Springer, Cham.

210. Schamp-Bjerede, T. (2012). What Clippy and Tux can teach us: incorporating affective aspects into pedagogy. *Högskolepedagogisk debatt*, (1), 47-56.

211. Schore, J. R., & Schore, A. N. (2008). Modern attachment theory: The central role of affect regulation in development and treatment. *Clinical social work journal*, 36(1), 9-20.

212. Schweikard, D. P. (2017). Cooperation and social obligations. *Distributed agency*, 233-242.

213. Selgelid, M. J. (2014). Moderate eugenics and human enhancement. *Medicine, Health Care and Philosophy*, 17(1), 3-12.

214. Sharkey, A. J. (2016). Should we welcome robot teachers?. *Ethics and Information Technology*, 18(4), 283-297.

215. Shim, J., & Arkin, R. C. (2017). An intervening ethical governor for a robot mediator in patient-caregiver relationships. In *A World with Robots* (pp. 77-91). Springer, Cham.

216. Shim, J., Arkin, R., & Pettinatti, M. (2017, May). An Intervening Ethical Governor for a robot mediator in patient-caregiver relationship: Implementation and Evaluation. In *2017 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 2936-2942). IEEE.

217. Shuttleworth, J. (2019). SAE Standards News: J3016 automated-driving graphic update. *SAE International.*

218. Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., ... & Chen, Y. (2017). Mastering the game of go without human knowledge. *Nature*, *550(7676)*, 354-359.

219. Simonite, T. (2019, February 13). Machines Learn a Biased View of Women. Wired. https://www.wired.com/story/machines-taught-by-photos-learn-a-sexist-view-of-women/

220. Skinner, B. F. (2014). *Contingencies of reinforcement: A theoretical analysis* (Vol. 3). BF Skinner Foundation.

221. Slim, F. (1999). Right here, right now. *Skint records* [prod.].

222. Sonkusare, S., Ahmedt-Aristizabal, D., Aburn, M. J., Nguyen, V. T., Pang, T., Frydman, S., ... & Guo, C. C. (2019). Detecting changes in facial temperature induced by a sudden auditory stimulus based on deep learning-assisted face tracking. *Scientific reports*, 9(1), 1-11.

223. Specker, J., Focquaert, F., Raus, K., Sterckx, S., & Schermer, M. (2014). The ethical desirability of moral bioenhancement: a review of reasons. *BMC Medical Ethics*, 15(1), 67.

224. Stroh, C. M. (2016). *Vigilance: The Problem of Sustained Attention*: *International Series of Monographs in Experimental Psychology* (Vol. 13). Elsevier.

225. Suikkanen, J., & Kauppinen, A. (Eds.). (2018). *Methodology and Moral Philosophy*. Routledge.

226. Surden, H. (2019). Artificial Intelligence and Law: An Overview. Georgia State University Law Review, 35.

227. Syrdal, D. S., Dautenhahn, K., Koay, K. L., & Ho, W. C. (2014). Views from within a narrative: Evaluating long-term human–robot interaction in a naturalistic environment using open-ended scenarios. *Cognitive computation*, 6(4), 741-759.

228. Taleb, N. N. (2007). *The black swan: The impact of the highly improbable* (Vol. 2). Random house.

229. Tamagnini, P., Krause, J., Dasgupta, A., & Bertini, E. (2017, May). Interpreting black-box classifiers using instance-level visual explanations. In *Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics* (pp. 1-6).

230. Terbeck, S., & Chesterman, L. P. (2014). Will there ever be a drug with no or negligible side effects? Evidence from neuroscience. *Neuroethics*, 7(2), 189-194.

231. Thai-Nghe, N., Drumond, L., Krohn-Grimberghe, A., & Schmidt-Thieme, L. (2010). Recommender system for predicting student performance. *Procedia Computer Science*, 1(2), 2811-2819.

232. The Editors of Encyclopaedia Britannica. (2020). Dionysus | Powers, Personality, Symbols, & Facts. Encyclopedia Britannica. https://www.britannica.com/topic/Dionysus

233. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and *Intelligent Systems*, Version 2. IEEE, 2017. http://standards. ieee.org/develop/indconn/ec/autonomous_ systems.html.

234. Thrall, J. H., Li, X., Li, Q., Cruz, C., Do, S., Dreyer, K., & Brink, J. (2018). Artificial intelligence and machine learning in radiology: opportunities, challenges, pitfalls, and criteria for success. Journal of the American College of Radiology, 15(3), 504-508.

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

236. Tonkens, R. (2012). Out of character: on the creation of virtuous machines. *Ethics and Information Technology*, 14(2), 137-149.

237. Topol, E. (2019). *Deep medicine: how artificial intelligence can make healthcare human again*. Hachette UK.

238. Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, 25(1), 44-56.

239. Transhumanist FAQ - Humanity+. Humanity+. (2020). Retrieved 28 August 2020, from https://humanityplus.org/philosophy/transhumanist-faq/.

240. Treanor, B., & Sweetman, B. (2016). Gabriel (-Honoré) Marcel. Stanford Encyclopedia of Philosophy.

241. Turchin, A., & Denkenberger, D. (2020). Classification of global catastrophic risks connected with artificial intelligence. *AI & SOCIETY*, 35(1), 147-163.

242. Tzafestas, S. G. (2016). *Roboethics. A navigating overview*. Heilberg: Springer.

243. Umematsu, T., Sano, A., Taylor, S., & Picard, R. W. (2019, May). Improving Students' Daily Life Stress Forecasting using LSTM Neural Networks. In *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)* (pp. 1-4). IEEE.

244. Vanderelst, D., & Winfield, A. (2018). An architecture for ethical robots inspired by the simulation theory of cognition. *Cognitive Systems Research*, 48, 56-66.

245. Vanderelst, D., & Winfield, A. (2018, December). The dark side of ethical robots. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 317-322).

246. Veloso, M., Biswas, J., Coltin, B., & Rosenthal, S. (2015, June). CoBots: Robust symbiotic autonomous mobile service robots. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

247. Verrugio, G., (2005). The Birth of Roboethics. In *IEEE International Conference on Robotics and Automation Workshop on Robo-Ethics*, 2005.

248. Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., ... & Oh, J. (2019). Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575(7782), 350-354.

249. Wagner, B. (2018). Algorithms and Human Rights: Study on the Human Rights Dimensions of Automated Data Processing Techniques and Possible Regulatory Implications. *DGI* (2017), 12.

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

251. Wallach, W., Allen, C., & Smit, I. (2008). Machine morality: bottom-up and top-down approaches for modelling human moral faculties. *Ai & Society*, 22(4), 565-582.

252. Wang, F. Y., Zhang, J. J., Zheng, X., Wang, X., Yuan, Y., Dai, X., ... & Yang, L. (2016). Where does AlphaGo go: From church-turing thesis to AlphaGo thesis and beyond. *IEEE/CAA Journal of Automatica Sinica*, 3(2), 113-120.

253. Warrick, P., & Homsi, M. N. (2017, September). Cardiac arrhythmia detection from ECG combining convolutional and long short-term memory networks. In *2017 Computing in Cardiology (CinC)* (pp. 1-4). IEEE.

254. Weaver, I. C., Cervoni, N., Champagne, F. A., D'Alessio, A. C., Sharma, S., Seckl, J. R., ... & Meaney, M. J. (2004). Epigenetic programming by maternal behavior. *Nature neuroscience*, 7(8), 847-854.

255. West, M., Kraut, R., & Ei Chew, H. (2019). I'd blush if I could: closing gender divides in digital skills through education.

256. Why New York City is getting an algorithms officer. (2019, November 20). CSNY. https://www.cityandstateny.com/articles/policy/technology/why-new-york-city-is-getting-an-algorithms-officer.html

257. Wikipedia contributors. (2020, June 15). Steve Woodmore. Wikipedia. https://en.wikipedia.org/wiki/Steve_Woodmore

258. Winfield, A. F., Michael, K., Pitt, J., & Evers, V. (2019). Machine ethics: the design and governance of ethical AI and autonomous systems. *Proceedings of the IEEE*, 107(3), 509-517.

259. Wiseman, H. (2014). SSRIs as moral enhancement interventions: a practical dead end. *AJOB Neuroscience*, 5(3), 21-30.

260. Wright, S. (2017). Virtue Responsibilism. In *The Oxford Handbook of Virtue*, Snow, N. E. (Ed.). Oxford University Press.

261. Yeung, K. (2018). A study of the implications of advanced digital technologies (including AI systems) for the concept of responsibility within a human rights framework. *MSI-AUT* (2018), 5.

262. Young, S. (2009). *Designer evolution: A transhumanist manifesto*. Prometheus Books.

263. Zagzebski, L. T. (1996). *Virtues of the mind: An inquiry into the nature of virtue and the ethical foundations of knowledge*. Cambridge University Press.

264. Zhang, Y., Weninger, F., Björn, S., & Picard, R. (2019). Holistic affect recognition using PaNDA: paralinguistic non-metric dimensional analysis. *IEEE Transactions on Affective Computing*.

Tomislav Miletić was born on May 7th 1984 in Rijeka. He has obtained a baccalaureate in philosophy from the Pontificial Gregorian University in Rome, Italy, and a masters of theology from the University of Zagreb, Catholic Faculty of Theology. In the period of writing this concluding this dissertation (2020/2021) Tomislav was an International Research Fellow (PE19056) of Japan Society for the Promotion of Science, supported by the "FY2019 JSPS Postdoctoral Fellowship for Research in Japan (Short-term)" in Kyoto HRI Laboratory, University of Kyoto, Japan. Tomislav works on ethical and symbiotic artificial intelligences.

Tomislav Miletić rođen je 7. svibnja 1984. u Rijeci. Ostvario je bakaleurat filozofije na Papinskom sveučilištu Gregoriana u Rimu, Italija, i magisterij teologije sa Katoličkog bogoslovnog fakulteta, Sveučilište u Zagrebu. U razdoblju zaključenja ove disertacije (2020/2021.) Tomislav je bio međunarodni istraživač (PE19056) Japanskog društva za promicanje znanosti, podržan od "FY2019 JSPS postdoktorskog stipendiranja za istraživanje u Japanu (kratkoročno)" u Kyoto HRI Laboratoriju, Sveučilište u Kyotu, Japan. Tomislav radi na etičkoj i simbiotskoj umjetnoj inteligenciji.

Published papers/Objavljeni radovi:

Miletic, T. (2015). Extraterrestrial artificial intelligences and humanity's cosmic future: Answering the Fermi paradox through the construction of a Bracewell-Von Neumann AGI, Journal of Evolution and Technology, Vol. 25:1, 2015, 56-73. http://jetpress.org/v25.1/miletic.htm

Miletic, T. (2015). Human Becoming: Cognitive and Moral Enhancement Inside the Imago Dei Narrative, Theology And Science, Vol. 13:4, 425-445. http://dx.doi.org/10.1080/14746700.2015.1082867

Miletic, T. (2020). Informed consent, military medical enhancement and autonomous AI systems: requirements, implications, concerns, in "Military and Humanitarian Medical Ethics" Volume "Ethics of Military Medical Innovation, Experimentation, and Enhancement", Editors D. Messelken/ D. Winkler, Springer International Publishing, 2020.

Miletic T., & Gilbert F. (2020). Does AI Brain Implant Compromise Agency? Examining Potential Harms of Brain-Computer Interfaces on Self Determination in: "Artificial Intelligence: a Multidisciplinary Perspective", Editor Steven S. Gouveia, Vernon Press.