AI Ethics Beyond the Anglo-Analytic Approach: Humanistic Contributions from Chinese Philosophy

*Paul D'AMBROSIO**

Abstract

That artificial intelligence (AI), algorithms, and related technologies could use a few good booster shots of "humanism" is widely apparent. In both program code and implementation, AI and algorithms have been accused of harbouring deep-seated flaws that conflict with human values. They are prime examples of the skew towards white, Western, men and demonstrate the bankruptcy in the face of neoliberalist, profit- and market-oriented social paradigms that this special issue seeks to address.

Currently, computer scientists and AI researchers who are looking to remedy these problems are often in favour of more data, more powerful machines, more complex algorithms—in short, that we should fix problems with AI by building better AI. In this view human beings and the world can be modelled in code—our lives, interactions, society, and our very selves can be broken down into data points which can be assessed by highly advanced technologies. When these scientists and researchers seek to broaden their approach they often look to philosophy. However, the philosophy they look to is overwhelmingly Anglo-analytic, which views the world in extremely similar ways. Both AI and Anglo-analytic philosophy argue for solutions to humanistic problems which are essentially mathematical. They share in seeing important concepts, such as persons, emotions, agency, and ethics, as mechanistic, atomistic, and calculable.

In this paper I will argue that Classical Chinese philosophy offer insightful resources for addressing the humanist problems in AI. Rather than arguing for mathematical solutions, or envisioning persons, emotions, agency, and ethics, as other rigid, atomistic, and mechanistic approaches, Chinese philosophy emphasizes transformation, interrelatedness, and correlative developments. Accordingly, it offers tools for appreciating the world, society, and ourselves as spontaneous, complex, and full of tension. AI can be programmed and used in ways that do not reduce the complexity and conflict in the world, but provide us instead with tools to make sense of it—tools that are humanistic in nature. To this end, Chinese philosophy can be a helpful collaborative partner.

Keywords: AI, algorithms, Chinese philosophy, humanism, machine learning, Confucianism, Daoism, comparative philosophy

* Paul D'AMBROSIO, Fellow of the Institute of Modern Chinese Thought, Philosophy Department, East China Normal University, Shanghai.
Email address: pauljdambrosio[at]hotmail.com

Etika umetne inteligence onkraj angloanalitičnega pristopa: humanistični prispevki kitajske filozofije

Izvleček

Da bi umetna inteligenca (UI), algoritmi in sorodne tehnologije potrebovali nekaj dobrih injekcij »humanizma«, je splošno znano. Tako samim programskim kodam kot tudi načinu uporabe se očita, da se v njihovi zasnovi skrivajo nekatere globoko zakoreninjene pomanjkljivosti, ki so v nasprotju s človeškimi vrednotami. So najboljši primeri tehnologije, ki služi belim zahodnim moškim, in s tem kažejo na bankrot vsakršnih naprednih načel, saj podpirajo razvoj neoliberalističnih, k dobičku in trgu usmerjenih družbenih paradigem, ki jih skuša obravnavati ta posebna številka.

Računalničarji in raziskovalke umetne inteligence, ki želijo odpraviti te težave, se trenutno pogosto zavzemajo za več podatkov, zmogljivejše stroje, kompleksnejše algoritme – menijo skratka, da bi morali težave z umetno inteligenco odpraviti tako, da bi ustvarili boljšo umetno inteligenco. V takšnem razumevanju naj bi bilo možno tako ljudi kot tudi ves svet modelirati v kodah; naša življenja, naše interakcije, družbo, v kateri živimo, in nas same je možno razdeliti na podatkovne enote, ki jih je z naprednimi tehnologijami mogoče vrednotiti. Toda ko te znanstvenice in raziskovalci želijo svoj pristop razširiti, se pogosto obrnejo na filozofijo. Vendar je filozofija, ki jo iščejo, večinoma angloanalitična, torej taka, ki na svet gleda na zelo podobne načine. Tako umetna inteligenca kot angloanalitična filozofija zagovarjata rešitve humanističnih problemov, ki so v bistvu matematične. Pomembne pojme, kot so osebe, čustva, delovanje in etika, vidita kot mehanistične, atomistične in izračunljive.

V tem prispevku bom pokazal, da klasična kitajska filozofija ponuja prodorne vire za reševanje humanističnih problemov na področju umetne inteligence. Namesto da bi zagovarjala matematične rešitve ali si osebe, čustva, delovanje in etiko predstavljala v luči togih, atomističnih in mehanističnih pristopov, kitajska filozofija poudarja preoblikovanje, medsebojno povezanost in korelativen razvoj. V skladu s tem ponuja orodja za vrednotenje sveta, družbe in nas samih kot spontanih in hkrati kompleksnih bitij, ki so polna napetosti. UI je mogoče programirati in uporabljati na načine, ki sicer sami po sebi ne zmanjšujejo zapletenosti in konfliktov v svetu, vendar nam lahko zagotovijo možnosti razumevanja le-teh. Tovrstne možnosti so že po svoji naravi humanistične. V ta namen je lahko kitajska filozofija koristna disciplina, ki nam lahko pomaga naučiti se sodelovanja s svetom in s soljudmi.

Ključne besede: konfucijanstvo, daoizem, primerjalna filozofija, kitajska filozofija, humanizem, strojno učenje

Introduction

Problems in AI ethics today are largely clustered around various biases, the enforcement of self-fulfilling prophesies, and goals prioritizing neoliberalist, profit- and market-oriented concerns. One of the earliest and to date most systematic studies of these issues is Cathy O'Neill's path-breaking *Weapons of Math Destruction* (2016). Summarizing these critical sentiments, the subtitle to her book reads: *How Big Data Increases Inequality and Threatens Democracy*. O'Neill's work calls for a reorientation towards more humanistic uses of data, algorithms, and the AI that relies on them. Dispelling the implicit assumption that because math and science are involved AI must be objective, she writes:

> Big Data processes codify the past. They do not invent the future. Doing that requires moral imagination, and that's something only humans can provide. We have to explicitly embed better values into our algorithms, creating Big Data models that follow our ethical lead. Sometimes that will mean putting fairness ahead of profit. (O'Neill 2016, 204)

Following O'Neill's lead many other data scientists have joined the call to "explicitly embed better values" into the programs that constitute AI, and various other algorithmic based tools being used on a daily basis. Research in widely disparate areas have come to the similar conclusions: technology largely mirrors the human thinking that creates, uses, and feeds it.[1]

For example, facial recognition software is notoriously skewed towards certain races, sexes, and genders.[2] Biases are also built into many other technologies in surprisingly simplistic ways.[3] Even seemingly objective AI tools, such as search engines, are deeply embedded with "data discrimination". As Safiya Umoja Noble (s.d.) shows, there is a "culture of racism and sexism in the way discoverability is created online". And "codification of the past" creates "feedback loops" which

1 "Feeds" here refers to big data, which is the "raw material" of nearly all AI today. Much of this big data is reliant, in one way or another, on observations of human behaviour. How exactly this is read, however, is to a large extent completely and necessarily determined by human programmers.

2 Some of the leading scholars in these areas include Joy Buolamwini (2019), Safiya Noble (2018), and Timnit Gebru (2020).

3 Lex Fridman relates a somewhat minor example, but one which shows just how prevalent and how blind many AI researchers and programmers can be. He describes working on technology to allow users to bypass password protections by identifying the unique way they pull their phones out of their pockets. Working on a team, Fridman and several colleagues were beginning to get satisfied with their progress when a female co-worker pointed out "many women don't have pockets, and carry their phones in their purse" (Fridman 2022, 104:15).

curate everything from social media to predictive policing.[4] Focusing on the latter, Sarah Byrne writes:

> Predictive algorithms hold up a mirror to the past, and project into the future. If historical biases and police practices inform where and whom the police surveilled in the past, they will also shape where and from whom cops collect the crime data that is fed into policing algorithms to generate those risk scores. It is a self-fulfilling statistical prophesy. (Byrne 2020, 16:30)

That much of the time, energy, and money is geared towards profit goes far beyond advertising tricks and ploys to attract attention.[5] Our entire economy has begun to shift towards what Shoshana Zuboff calls "surveillance capitalism", which mines "human experience as free raw materials for translation into behavioral data" (Zuboff 2018, 8).[6]

Dealing with these issues would make AI more "humanistic"—if by this we mean better aligning technology with our values. This requires thinking about AI and technology in somewhat unfamiliar ways, i.e. looking outside of ways AI itself currently operates, or at the very least changing our relationship to this technology. Otherwise, we risk merely patching up a few problems or fixing some things while messing others up. Indeed, this is already implied in what O'Neill, Noble, and Byrne identify as the main issue: further drawing on resources which reflect the method and content we already use might only lead to further codification of the very biases and practices we are trying to avoid.

The dominant approach in addressing AI-related issues today, and one that more or less mirrors how AI programs are already constructed, is Anglo-analytic philosophy[7]. Methodologically, as well as in terms of basic assumptions about key concepts

4 Cathy O'Neill also investigates the algorithms used to inform recidivism assessments. These models, O'Neill demonstrates, "codify prejudices and penalize the poor" (O'Neill 2016, 210). Moreover, she along with many others have definitively shown that biases are rampant in credit card companies, insurance rates, *résumé sorting, mortgage assessments,* or basically anytime any algorithm is used. Brian Christian's *Alignment Problem* (2020) also discusses many of these issues in detail.

5 Tim Wu's *The Attention Merchants* (2016) and Joseph Turow's *The Aisles Have Eyes* (2017) give excellent discussions of these issues.

6 Or, as Wikipedia describes: "Surveillance capitalism is an economic system centered around the capture and commodification of personal data for the core purpose of profit-making," (https://en.wikipedia.org/wiki/Surveillance_capitalism (accessed November 21, 2022)).

7 The description of Anglo-analytic philosophy given here does not claim to represent the entire tradition—nor would that ever be possible. What is of interest is "Anglo-analytic philosophy" in two general senses: first, the type of philosophical resources utilized by people who develop or theorize about AI, and second, the general approach of Anglo-analytic philosophy. Much of this

including persons, emotions, agency, and ethics as well as understandings of meaning, AI and Anglo-analytic philosophy are quite similar.[8] Unsurprisingly, communication between AI researchers and Anglo-analytic philosophers is extremely easy and common—and that is because the not-so-distant cousins are speaking dialects of the same language. However, Chinese philosophy provides a great resource on both counts. It runs counter to the way AI and Anglo-analytic philosophy thinks about persons, emotions, agency, ethics and meaning, and methodologically, what counts in Chinese philosophy (e.g. how arguments are made) and even what issues are discussed, is quite different from Anglo-analytic philosophy as well. But perhaps most importantly, already by looking to something as different as classical Chinese philosophy, AI research would necessarily embody an orientation toward reflecting on the unfamiliar in unaccustomed ways. This is precisely what is needed.[9]

This paper thereby seeks to explore ways in which Chinese philosophical methods and concepts could contribute to rethinking AI research, programming, and implementation in more humanistic ways. The paper will be broken down into four main parts. In the first section ("Methods") we will compare some of the differences between the similar approaches shared in Anglo-analytic philosophy and AI with the Chinese tradition. The second section ("Technology and Tradition") will examine the call for humanism in the *Age of AI*, and consider how the tradition-focused method of Chinese philosophy is well equipped to meet these challenges. The concentration in these first two sections will be on methodology. Section three ("Concepts") will

paper concerns how Anglo-analytic philosophy as a way of thinking is markedly different from Chinese philosophy—which again is outlined in extremely broad terms. Additionally, while the purpose of this paper is to reflect on how insights from Chinese philosophy might be useful, this by no means rejects the importance of the Anglo-analytic approach, which is certainly quite useful for AI. Another way to frame this might be "a certain type of Anglo-analytic approach"—which is represented by some of the thinkers discussed here, but more broadly is found in the literature on AI and related references. This article does not attempt to give a summary of that literature, but address its general orientation. There are a vast number of limitations here, and perhaps those who work in the Anglo-analytic tradition would say I am fighting a straw man, but many other readers might find this very useful. I can only acknowledge the limitations and hope that this article proves useful to at least some of the readers.

8 The major point in this paper is that AI researchers draw on Anglo-analytic philosophy in certain ways, and I seek to discuss that. I am not so much discussing Anglo-analytic philosophy as such. I do make certain generalizations about it, but those are meant to be useful for understanding how AI researchers have approached issues, not how Anglo-analytic philosophy has.

9 Though Chinese philosophy is not to be put on too high a pedestal and we should readily admit that other traditions, including the Western one, could provide good resources for reflecting on ways to make AI more humanistic. Moreover, in content and method Chinese philosophy has close relatives all over Europe and America—and they are quite different from the Anglo-analytic approach as well. This paper is not arguing that Chinese philosophy is a silver bullet for all, or even any, of our problems with AI. But important contributions can certainly be made if AI researchers worked with specialists of the Chinese tradition.

then turn to specific concepts: persons, emotions, agency, and ethics. Here we will again note general similarities between the way these concepts are thought of in Anglo-analytic philosophy[10] and AI while comparing the relatively unique ways they are thought of in Chinese philosophy. The fourth section ("Math and Meaning") will explore the contemporary discourse that argues humans should model AI "thinking" and rely even more heavily on numeric models and other mathematical structures. This will be contrasted with a philosophical appreciation of complexity and understandings of *meaning* fundamentally distinct from math—early Chinese philosophy is a good example of this approach.[11]

Methods

The neuroscientist Andrew Huberman has suggested understanding addiction as "a progressive narrowing of the things that bring you pleasure". Conversely, happiness, or, as he ventures to say, "a good life" is the "progressive expansion of the things that bring you pleasure" (Huberman 2022, 6:50). There is an interesting corollary here when we look at the general orientation of Anglo-analytic philosophical approaches (including Anglo-Chinese/comparative philosophy) in comparison with the Chinese tradition. The former often tend to seek to hone in on ever more narrow aspects of greater issues, while the latter is relatively expansive in what it takes into account. This section explores some differences between the Chinese tradition and contemporary Anglo-analytic philosophy,[12] relying heavily on general distinctions in orientations. These can be particularly useful for the current project—especially as "AI"[13] and "Chinese

10 Again, I do not mean all of Anglo-analytic philosophy, but certain trends related to it, especially those which emphasize "technological reasoning", "mechanical appreciations", "abstract consideration of things as distinct from their environments", and "algorithmic ideology".

11 Given the breadth of the topic, this paper has many limitations. Chinese philosophy, Anglo-analytic philosophy, and AI are far more complex than can be presented, and there are counter examples to every generalization made. The purpose of this paper is to inspire certain types of reflection, and by no means to provide an exhaustive account of debates and issues raised here. For those who find this description of Anglo-analytic philosophy problematic, please replace it with "some technical and mechanistic ways of approaching philosophical issues".

12 Much of what is said about Anglo-analytic philosophy applies to Anglo-Chinese/comparative philosophy as well.

13 In this paper AI is often used as a gloss for algorithms, machine learning, and deep learning—and each of these terms can be further broken down. While non-experts may find this quite acceptable, it should be noted that the difference between these terms is significant not only for programmers and technologists, as there are huge philosophical implications as well. However, this has not been (to my knowledge) discussed in detail.

philosophy"[14] are already huge generalizations, and combining them necessitates using broad strokes.

One general orientation in Western philosophy, pioneered by Socrates and Plato, is the distinction between "essence" or "substance" and its "attributes". The thing-in-itself is what counts, and those aspects which are accidental—which can include anything from size and colour to emotions and social relationships—should be purposively neglected to whatever extent possible. Here thinking is understood as an ascent towards this "true essence", and moulded into the shape of a pyramid—as one progresses, one moves higher, as one treads on increasingly narrow space, the expansive concrete world gets further away. For example, when thinking with Euthyphro about piety and justice, Socrates characteristically rejects concrete examples in searching for a universal definition. As their dialogue moves along, it becomes increasingly narrow and abstract. Interestingly, neither Socrates nor Euthyphro consider social roles or emotions as integral for their discussion of "ethics".

In the *Analects* we find the opposite orientation. Confucius is well-known for refusing to provide the abstract definitions his students consistently ask for, answering instead with the very type of concrete examples Socrates rejects. When considering whether or not a son should cover for his father who has stolen a sheep, Confucius prioritizes social relationships and emotional considerations. One could imagine the Chinese sage chastising Euthyphro—not for not knowing the universal definition of piety, but for simply being a bad son. Rather than appreciating the complexity and difficulty of his situation, Euthyphro is encouraged to look at matters so abstractly that actions and concepts are all but completely isolated from their context. In the *Analects* we find phrases that celebrate the opposite: "There is nothing I must do and thinking I must not do" (*Analects*, 18.8), or "ornamental aspects are like basic dispositional aspects and basic dispositional aspects are like ornamental aspects" (*Analects*, 12.8) point to the centrality of concrete particulars in ethical reflections.

The general differences here can be summarized as follows: With Socrates we seek to exclude concrete particulars in order to grasp essences and come up with universal definitions. We narrow what counts, we abstract from environments, and evaluate without recourse to models or examples. All "ornamental aspects" are completely strained out so we are left with only "basic dispositional aspects". Further, we must be completely responsible for our own reasoning. With Confucius,

14 We could start, of course, by simply asking whether or not there is in fact any "Chinese philosophy". Many other disputes about labelling or categorizing Chinese thought exist and are alive and well (some more or less just born) as will be outlined below.

on the other hand, we seek to gather as many particulars as possible, we do not divorce basic dispositions (which function similar to "essence") from attributes, and reject universal definitions favouring instead examples, exemplars, and models. There is a progressive expansion of what matters, and we aim to appreciate as many environmental factors as we reasonably can. We are not completely responsible for our evaluations, we view ourselves and our own agency as part of a tradition and interrelated with everything else.

In broad terms, the development of these two traditions carried on along these respective lines. Western thought emphasized essences, the rejection of particulars, and sought universal definitions through relatively abstract and concept-based discussions. Moreover, many great Western philosophers began their work with criticisms of those who came before, and proclaimed to have a better handle on the "Truth" than their own teachers. The Chinese tradition was "tradition-focused"—philosophy developed mainly through commentaries which started with the classics and claimed to merely be reinterpreting them to resonant with current socio-political situations. Contextualizing with an emphasis on practical issues was commonplace. So too was modelling the past. Learning from, rather than overcoming, was the attitude toward predecessors. (We may note that Western thinkers often assumed they were more original than they really were, while Chinese commentators often claimed to be less original than they really were.)

Anglo-analytic philosophy can be seen as the development of Western philosophical thought with a particular "linguistic turn". Here hypothetical questions—which even to this day are all but absent in Chinese philosophy[15]—dominate. Analogously, nearly every term or concept is narrowed as much as possible. For instance, the analytic tradition does not admit "lying" as a topic; one must differentiate between at least several types of lying. Unsurprisingly, culture, tradition, differences between languages, even the fact that a person changes over time, or that they interact differently depending on their interlocutor or environment, are all challenges for this methodology. It reduces these aspects, just as it does "lying" or anything else, to linguistic math problems. In this way we hope to trade only with absolutes, with universals, and, of course, Truth with a capital T as the main targets.

Herein lies a core difficulty with the relationship between philosophy and AI. When those who seek to inject more "humanism" or humanistic values—including morality and ethics—or even to simply reflect on assumptions made by the programmers and programs themselves (as well as their implementation) turn to

15 By "Chinese philosophy" I often mean Chinese philosophy in China and Chinese, not the study of Chinese philosophy done in English.

Anglo-analytic philosophy, they often look in an only slightly unfamiliar mirror. There is a level of convenience which perpetuates this relationship and sterilizes it to some extent. Both AI and Anglo-analytic philosophy are dialects of the same language. Conceptually, they are akin as well—they tend to narrow the world through reducing cultural differences, ignoring tradition, pretending simplistic linguistic equivalences between languages, and treating people as rational agents with an agency which is static, singular, and individualistic. Thus, when trying to improve the current problems with AI, looking to Anglo-analytic philosophy can certainly be fruitful,[16] but looking beyond this approach may result in even more unexpected contributions.

"Inspiration" captures perhaps the single greatest difference between traditional Chinese philosophy (and traditional Western philosophy) and the approach of Anglo-analytic philosophy and AI research. Ignoring how we might generalize various forms of inspiration in the East and West, we can begin by noting that Chinese philosophy is often described as "practical". The discussion on narrowing versus expanding above has often taken shape as "Western philosophy is abstract and universal, Chinese philosophy is concrete and practical". Some have sought to claim that Chinese philosophy is not merely practical, as it has a lot to say about metaphysics, ontology, and the like. For example, in the hands of Wang Bi 王弼 (d. 249) the *Laozi* 老子 (*Book of Master Lao*) becomes a treatise on "nothingness" (*wu* 無) as an ontological source of all things, and the metaphysical basis for their functioning. Truly, Wang Bi also explores the socio-political implications. Applying his "root-branches" model Wang says the ruler should model the nothingness of *Dao* as the root, and allow the people as branches to become full—and this will be the concrete implementation of non-action (*wuwei* 無為) and self-so (*ziran* 自然). But Wang Bi's commentary offers much more as well. The "practical" aspects of what he says might also be understood as "practical" for the individual, too. Fully adopting the ideas of non-action and self-so himself, Wang completely refrains from telling the individual what they should or should not do in a precise and mechanistic manner. Rather, Wang's work is practical for readers in the sense of inspiring philosophical reflection—in a word, Wang Bi *inspires*.

Contemporary academia has been moving toward greater measurability and quantification in nearly all areas. Academic journals today often ask reviewers to evaluate papers based on tables with Likert-style rankings from 1 to 5. The exact questions they ask are a "feedback loop" or mirror of the very measurability and qualification this structure promotes, with questions such as "How likely is this

16 I am by no means suggesting that analytic philosophy has not nor cannot make significant contributions to AI. For example, Paul Grice's work on the extended meaning of utterances has been appreciated by computer scientists with great effect (see Russell 2019).

paper to be cited by other researchers?" No academic journal I know of asks about inspiration (as if this could be measured)—though the texts we work on are all incredibly inspiring, and thus, one assumes, that is why we work on them in the first place. And their lack of providing mathematical solutions is what sparks debate along with inspiration. Reviewers are also not asked about whether the article resonances with or even is aware of traditional scholarship. Even in Chinese philosophy journals demonstration of competence in language, classic commentaries, and cultural sensitivity is not assessed.[17] We may also say—though again there are plenty of exceptions—that Anglo-analytic philosophy (and AI research in some ways) does not prioritize inspiration, either. Academics are famously bad at communicating with non-academics, but it is no stretch to suggest that Anglo-analytic philosophers have a particularly difficult time in inspiring the public, or even in convincing others about the meaning (in either sense of the word) of their work.

While sound arguments with logical reasoning and a sharp eye for the Truth provide a standard in Anglo-analytic philosophy, one of the key markers for "good" philosophical research in Chinese thought is the rather obtuse notion of *tong* 通, which literally means "without obstruction", "open", "unimpeded" or "through". It is used, for example, when making sure a road is not a deadend: "Is this road *tong*? (這路通不通？)" In philosophical arenas it means something like "resonance" or "thoroughness". A philosophy student is extremely happy to hear from an advisor "your work is *tong*", as this means that the work resonates well with whatever text they are working on, and that it fits well with certain streams of traditional scholarship. We might conceptualize it like this: Whatever idea is *tong* runs through a text and parts of the tradition. It communicates well and is not in direct conflict with other important aspects of that text or the tradition. Accordingly, the "argument" and "reasoning" utilized to achieve *tong* relies on the very philosophical approach already found in the *Analects*, gathering as many particulars (other parts of the text and tradition) as possible, and heavily referencing models and examples—often in the form of extensive quotations in what might be called "appeals to authority", but are better understood as "humbly learning from masters".

Chinese philosophy can thereby be understood as a "tradition-focused" philosophy. There are two related connotations of "tradition-focused". First, it means learning from and referring tradition—much of what *tong* is all about. Second, it means not taking persons, concepts, or thinking out of their traditional context. Importantly, this is not limited to dealing with the past. Today tradition also

17 The most extreme version of this is evidenced by those who proclaim that Western academia is "racist", and yet only cite other Western academics in their work. (Moreover, it is quite obvious that there is far more overt racism and sexism in Chinese universities, which these scholars similarly idealize as more diverse and pluralistic than, for instance, North American ones.)

needs to be prioritized in the way we think about other people we encounter, about ideas people have, the way they think, and much else. As we will explore in the following sections, a tradition-focused philosophy cares deeply about culture and language, it does not view things in isolation, it emphasizes the importance of relationships and environments, and takes inspiration and meaning as final evaluative standards.

Technology and Tradition

In his work on algorithms and digital dilemmas Roberto Simanowski asks: Are we "on the brink of a society that views social, political, and ethical challenges as technological problems that can be fixed with the right algorithm, the best data, or the fastest computer" (Simanowski 2018, 209) or more advanced AI? In much of the world today the answer is put into practice in implicit and explicit ways every day. And it trends strongly towards the affirmative.

One such example of the explicit push for this type of thinking is found in *Algorithms to Live By: The Computer Science of Human Decisions* (2016), which says enough already in the title. The "interdisciplinary" nature of the work further drives home the point. It is co-authored by Brian Christian, who researches computer science and Anglo-analytic philosophy, and Tom Griffiths, who works on psychology, cognitive science, and machine learning. While their interests are broad, their approach is fundamentally the same. Be it computer science and Anglo-analytic philosophy or cognitive science and machine learning, the way these authors view the world relies on the same foundational assumptions about "social, political, and ethical challenges as technological problems that can be fixed with the right algorithm, the best data, or the fastest computer or more advanced AI". The two claim that the algorithms that comprise AI systems are not just for computers, they actually have a lot to teach humans about the best way to live. Already in the first chapter, titled "Optimal Stopping" the authors claim that the algorithm for the 37% rule can be used to find anything from the optimal apartment or parking place to "optimal" spouse.

The 37% rule says that if you have to make a decision about choosing something (or someone) first set a predetermined amount of time you allot for looking for that thing, then spend the first 37% of that time just looking—do not make any choices—then after that 37% of time has passed pick the first candidate you like (as compared with the first 37%). If you assume that you cannot go back and choose one of those in the first 37%, and that any choice you make after the first 37% will be successful (e.g. she will say "yes") then mathematically, this is the best

strategy. The problem with the 37% rule is not that it does not align with our values nor that it ignores the idea of romantic love—though arguments can certainly be made along these lines—rather, the problem is that is asks us to conceive of the world using a model that simply is not accurate, and further it encourages a type of thinking that has little to do with the actual world. For example, in what (meaningful) situations can you reasonably designate a predetermined amount of time to make important decisions? The authors reference marriage, but setting specific time limits is an odd and often impractical way to think about finding a spouse (not to mention the subject themselves will change over time). Additionally, there are very few situations in which one cannot go back to previous options, and there is no guarantee in that one's choice after a certain period will be accepted. (We might also think, if everyone accepted the 37% rule in dating this would completely change how people thought, acted, and made decisions about dating. Knowing that oneself or one's date was still within the confines of the first 37% would certainly have an impact, as would being aware that the 37% was already reached. It just creates another self-fulfilling prophesy.)

The idea of using the 37% rule in real life is bizarre, if not comical. More importantly, it says absolutely nothing about *how* the choice is made. What qualities should we look for in an apartment or spouse? How do we evaluate them? What should our standards be? How do we balance them with other factors? These and a whole host of other questions, which are exceedingly more important, are completely ignored by the 37% rule. Mathematics might give us some interesting notes about how to spend time, but it says nothing meaningful about our actual interactions with the world.

Interestingly, Christian and Griffiths open their "Optimal Stopping" chapter with an epigraph quoting Jane Austen: "If you prefer Mr. Martin to every other person; if you think him the most agreeable man you have ever been in company with, why should you hesitate?" They could hardly have chosen an author who is more opposed to the mechanistic and technocratic "optimal stopping" approach to love. The Jane Austen quote seems to work. But only because it is taken out of context. And this is true of the entire approach the 37% rule promotes. Without context, i.e. without the real world in all of its complexity, the 37% rule might just be the best way to think about decision making. We do, however, live in the world where one's self, interactions, and the world itself cannot simply be reduced to math problems.[18] Out of context, Jane Austen's quote can be used to promote "optimal

18 The authors do admit that we do not live in conditions that make the 37% rule as mathematically sound as it might appear to be. However, the entire book is premised on steering human thinking and decision making towards these types of algorithms which do view the world in a very narrow way. Or, as they put it, the 37% rule is still "the best possible strategy" (Christian and Griffiths 2016, 14).

stopping". However, this quote comes actually from Emma criticizing Harriot for asking her for advice in matters of her own feelings. The sentence directly preceding the one quoted in the epigraph reads: "'Not for the world,' said Emma, smiling graciously, 'would I advise you either way [i.e. to marry Mr. Martin or not]. You must be the best judge of your own happiness.'" Emma no more seeks to have the world than to chalk it up to just so many math problems and data points. The heroines of Austen's novels are not just against traditional approaches to marriage, they are markedly against calculations or "algorithmic thinking" in general.

Simanowski has serious worries about the type of ethos promoted by the technocratic and mechanistic orientation Christian and Griffiths celebrate, and comes down on this thinking much harsher than Emma or Austen. For example, apply this type of thinking to education (which many parts of the world already do) and students become focused on competence and test scores rather than learning. They excel at taking tests but are hardly reflective. Or as Martha Nussbaum criticizes, "nations all over the world will soon be producing generations of useful machines, rather than complete citizens who can think for themselves, criticize tradition, and understand the significance of another person's sufferings and achievements" (Nussbaum 2010, 2). Christian embodies exactly these fears of Simanowski and Nussbaum, when he says, "There is a very deep resonance indeed between some of these ideas in computer science [regarding the best algorithms for machine learning] and the same fundamental algorithms of learning that evolution found" (Christian 2022, 22:40). In other words, humans and computers use the same methods for learning—and this rings true when we think of human education in terms of "objective functions" and other standards of computer science. Or what Nussbaum calls "useful machines" and Simanowski describes as reducing learning to obeying traffic rules. Changing our perspective to appreciate inspiration, meaning, and tradition we are forced to think quite differently about mapping algorithms onto human thinking. Indeed, there are many cases where this simply will not meet our real-world demands. In *The Age of AI* (2021), Henry Kissinger, Eric Schmidt, and Daniel Huttenlocher concentrate exactly on those types of issues—which they argue are the most pressing.

According to Kissinger, Schmidt, and Huttenlocher nothing less than a "new guiding philosophy" (Kissinger et al. 2021, 224) is required for our age of AI. They sharply disagree with the explicit widespread agreement in AI research that "everything can be figured out". Accordingly, as technology advances at unprecedented rates, at nearly unimaginable scales, and with unpredictable results, our solutions to the problems that arise is not ever more progress—we need deliberate philosophical reflection. They compare our current state to the Enlightenment, where "new technology engendered new philosophical insights, which, in turn,

were spread by the technology" (ibid.). Again, this is not a problem to be solved by mathematicians, or those whose thinking is characterized by ahistorical, linguistically uniform, and narrowing tendencies. Translation provides a provocative example. As relations between nations is increasingly mediated through AI translation, the potential for error, and for false-confidence—for example in an erroneous AI translation of sensitive information—intensifies. As Kissinger, Schmidt, and Huttenlocher note: "AI translators will facilitate speech, uninsulated by the tempering effect of the cultural familiarity that comes with linguistic study" (ibid., 222). Our philosophy of AI needs to be highly sensitive to the types of analysis, thinking, and imagining about the world that mathematical models are blind to. The cultural implications of AI, the role of humans in their collaborative future with AI, and the types of values, norms, and ethics we want to express in and through AI require sensitivity to the particularness of the human condition, and its "spontaneous experience of reality, in all its contradiction and complexity" (ibid., 219).

The "tradition-focused" aspects of Chinese philosophy are well suited to meet the challenges Kissinger, Schmidt, and Huttenlocher raise. Admittedly, we may very well need something new, a philosophical revolution akin to the Enlightenment, but it needs to be one that is critically sensitive to culture, language, and tradition, and does not see humans, their interactions, or the world as reducible to just so many mathematical algorithms. If the world really is spontaneous, complex, and full of contradictions, then Anglo-American philosophy is an ill-suited tool, and AI researchers need to expand their resources for humanistic reflection. As will be further demonstrated in the following section, Chinese thought accepts and works directly with spontaneity, complexity, and contradiction. Drawing on Chinese philosophy as a uniquely suitable resource to move forward into our collaborative future with AI is therefore a viable option, and one that should be taken seriously for other reasons as well.

In addition to offering a useful methodology to tackle the problems we face looking at our age of AI and the suitability of its major concepts (which will be discussed below), there are political reasons to take Chinese thought seriously. Kissinger, Schmidt, and Huttenlocher do not tire of pointing out that the US *and* China are the front runners in the age of AI. As we come to terms with this, we are certainly going to need to understand Chinese thought better. The Chinese tradition needs to be taken seriously for, if not moral, ethical, or cultural considerations, at least for diplomatic reasons. This will not mean, as some have suggested, simply plugging in codes for "filial piety" and "using chopsticks" next to "justice" and "forks". The entire way we approach AI will need to be constructively reworked. Specifically, the views on persons, emotions, agency, and ethics

as well as understandings of meaning, are in a dialectical relationship with the tradition-focused methodology, and understood in ways markedly different from the conceptions held by many AI researchers and Anglo-analytic thinkers.[19]

Concepts

The affirmation of ceaseless change marks an underlying assumption in Chinese thought. Conversely, Anglo-analytic philosophy seeks to overcome change to whatever extent possible. Analysis in the latter approach prefers discrete, abstract, and isolated subjects. When change is admitted, it is given as simplistic a rendering as possible. Chinese thought ventures to the other extreme. Here change is not only affirmed, but its presence can be taken as a point of emphasis. With its tendency to bring more and more variables into reflective consideration, "change" is often discussed as "transformation"—and through this concept we can get a better grasp of the general approaches to people, emotions, agency, ethics, meaning, and contingency in Chinese thought.

While we may speak of transformations as "occurring", it is more accurate to understand everything as constantly transforming. We normally only notice larger changes (hence we say "change occurs"), but we all know that smaller changes are happening all the time. To borrow from Guo Xiang 郭象 (d. 312) we can say that everything is some grouping of natural dispositions (*xing* 性). These dispositions might have certain innate tendencies, but whether or not they move in those directions—and whatever directions they do end up moving in—is all a result of their transforming with all other natural dispositions. All natural tendencies are ceaselessly transforming in their interactions with all others. These interactions are not "secondary", they do not happen, for example, when you go to the store but not when you stay home. Everything is always interacting with everything else in its environment—regardless of whether that environment is the most sterile room on earth or the deep Amazonian rain forest.

Of course tracing every interaction is impossible. Not only are there simply too many, but they are simply too complex as well. The web of everything transforming as it interacts with everything else means that when one attempts to isolate some interactions they are only viewing one small dimension of what is really going on. (Modern biologists and physicists also admit this point—and some even chastise analytic philosophy for borrowing from modern science in nearly

19 Specific discussions of theoretical differences between Chinese and Western approaches to AI are still being developed, for some insights in more practical application and regulation see (Roberts et al. 2021).

all regards except this one.)[20] One informative way to think about the differences between Chinese philosophers is in terms of where they demarcate the lines of what should or should not be included in their reflections.

Conceptions of the person which appreciate constant transforming with the environment view social interactions and relationality as key components—we may even say that there is no meaningful way to describe a person outside of social roles and relationships. We know well that our bodies are constantly interacting and changing with the environment, and it is no large imaginative leap to realize this in terms of personal identity and meaning. Whether it is language, likes, and hobbies, or religious, political, and philosophical convictions, everything that meaningfully gives people a sense of self, and a sense of others, comes from their environment—and specifically, what they learn from others and how these interactions play out. In early Chinese philosophy people are not thought of as abstract individuals whose social relations are accidental. Indeed, it is precisely the contrary, outside of social relations and interactions with their environment there is no meaningful way to speak of anyone.

Summarizing the way programmers and much of society view people, Shoshana Zuboff describes a type of "individualism" which "shifts all responsibility for success or failure to a mythical, atomized, isolated individual, doomed to a life of perpetual competition and disconnected from relationships, community, and society" (Zuboff 2018, 33). People are thought of as rational agents who are not constituted by their environment and relations, but rather can stand apart from them and deliberate about exactly which parts will be, and how they will be, incorporated into themselves. This break down of people, interactions, and environments into discrete components that can be tinkered with mechanically, is echoed in views of agency as well. Someone might have the desire to do X, but they can rationally reflect—outside, somehow, the influence of that desire—and decide to do Y if they so please. Like relationships, community, and society, a person's emotions, feelings, and desires can be incorporated into decision making, and the way a person understands themselves and others. But neither the "external" relationships, community, and society nor the "internal" emotions, feelings, and desires meaningfully comprise who a person is in an inalienable manner. (Henry Rosemont would apply this specifically to analytic philosophers, too.)

The early Chinese view on emotions could hardly be more different. The very word for "emotion" is *qing* 情, which, tellingly, no more connotes an individual's emotion than the environment. In other words, there is a thorough recognition that one's emotions are intimately tied to the environment, and that environments

20 This will be discussed in more detail below.

elicit certain emotions.[21] Similarly, there is no hard and fast distinction between emotions and reason. Again, terminology helps demonstrate the point: *xin* 心 refers to both the house of thinking and feeling, and is thereby nearly universally understood as some variant of "heart-mind". Truly, it would be incredibly difficult to find terms that would even allow for a distinction between heart and mind, or "reasoning about something" as opposed to "feeling about something" in early Chinese thought.[22]

A brief glimpse into the specifics demonstrates this point: One of the major goals of early Confucianism is to provide guidance for people to become moral or ethical members of society. The strategies for cultivating moral persons do include rational arguments that aim to convince people. But more often, and more importantly, we find a host of discussions aimed at emotionally inspiring people to do what is moral or "humane" (*ren* 仁). Why should one bury one's dead parents? Just look at unburied people (*Mencius* 3A:5). Or, to rephrase *Analects* 17.21: What if I don't want to mourn for three years after my parents die? Well, good people don't enjoy parties for a while after their parents die, you should at least try refraining from them and see if this doesn't help you have more appropriate thoughts and feelings. And that's it, no further "argument" is needed. In this way people are sometimes cultivated with emotion and sometimes with reason leading the way. Regardless, there is an expectation that when one leads, the other will eventually follow.[23]

Thinking in this way entails a very particular view of agency. The person does not direct their actions from a purely rational capacity that sits somehow outside the ultimate influence of environments, and even one's own feelings and desires. Agency is seen instead as completely tied up in emotions and environmental factors. Whatever one ultimately decides to do cannot be isolated into pure reason or some abstract and isolated decision-making process—it is always thoroughly effectuated by the whole person: their thoughts and feelings, their interactions, upbringing, tradition, language, and many other factors. In this way agency itself is subject to cultivation as well. When a person is young they are not good at critically reflecting on their own thoughts, emotions, desires, and interactions or

21 Again, this is a simple generalization. One of the best examples of an early Chinese thinker who vehemently disagrees is Ji Kang 嵇康 (d. 262), who famously argues that music does not necessarily elicit or come from particular emotions. For a discussion of this essay see Rošker (2014).

22 In modern Chinese we can say "我觉得" which means "I think/feel …" in certain contexts, there is more of an emphasis on thinking or feeling. And while we can separate the two, the phrase indicates a mixture where thinking and feeling should not be cleanly delineated.

23 On contemporary reiteration of this point has been popularized by Johnathan Haidt. See, for example, his work on "moral emotions" (Haidt 2003).

surroundings. In contrast, if a person cultivates themselves well then their ability to critically reflect becomes greatly enhanced.

In Confucian thought the best ways to develop one's agency is to make sure one has good influences.[24] Studying poetry, history, and other classic texts is a good way to learn from exemplars, so too is surrounding oneself with good people—particularly one's friends and teachers. But there is also a strong concern for one's neighbourhood, as it is recognized that this will be quite influential as well. Likewise, bad people are to be avoided.[25] And people are told to do good things even if they do not completely think or feel them. The practice of ritual is a practice of cultivating one's thoughts, feelings, and agency through prescribed bodily behaviours. For example, one should mourn for three years after one's parents die. Even if one wants to party, or especially if one wants to party, one should refrain from doing so in order to become a better person, i.e. a person with good thoughts, feelings, and agency. A modern-day example is making children say "thank you". Even if they do not feel appreciation, and in fact especially if this is the case, parents will force them to say "thank you". The hope is that the children will eventually develop a sense of gratitude. (We can apply this to "sorry", as well.)

In opposition to the Chinese conception of an agency that grows, many Anglo-analytic philosophers and AI researchers take agency as a ready-made power. Moreover, this power is critically divorced from external (non-rational) influence. Even Zuboff, who comes down so harshly on the "mythical, atomized, isolated individual, doomed to a life of perpetual competition and disconnected from relationships, community, and society" that dominates AI research, still speaks of the "sovereignty of the individual" (Zuboff 2018, 6) and holds "self-determination" (ibid., 18) as an ideal. Her work on surveillance capitalism is largely organized around demonstrating that AI is able to "nudge, tune, herd, manipulate, and modify behaviour in specific directions by executing [even subtle] actions" (ibid., 200). We must resist the malicious AI utilized in surveillance capitalism because it tampers with the sovereignty of our autonomous power. In other words, even while working against perspectives of the individual as atomized and isolated, Zuboff views the person in much the same way, and is especially focused on a

24 This is not a vicious circle, one has some inclinations toward what is good, and one needs to develop them. More importantly, however, Chinese thought does not worry about abstract logical arguments when it comes to recognizing basic differences between good and bad. See for example Robber Zhi sections in the *Zhuangzi* 庄子 and the famous line "robbers also have a way (盗亦有道)." (*Zhuangzi* 10.1)

25 Confucius said, "Thinking of what is good [and pursuing it] as if it could not be reached; thinking of what is not good [and avoiding it] as if it is boiling water (孔子曰：「見善如不及，見不善如探湯。)." (*Analects* 16.11).

type of agency that is all but entirely isolated from environmental factors. A more expansive analysis would recognize that our agency is always being influenced by other factors, and especially our emotions, which must always be tied to our environment. Rather than criticizing AI for "herding" or "manipulating" people, we should note that it influences them in ways or with methods that we do not approve of.[26] The influence itself is inescapable—decision-making is necessarily influenced by environmental factors.

As the biologist/neurologist Robert Sapolsky notes, when trying to understand agency we must recognize the influence of many factors, including, for example "the sensory environment you were in the previous minute, the hormone levels in your bloodstream that morning, whether you had a wonderful or stressful last three months and what sort of neuroplasticity happened, what hormone levels you were exposed to as a fetus, what culture your ancestors came up with and thus how you were parented when you were a kid" (Sapolsky 2021, 48:45). We can at least start with these elements, though there are many more. Sapolsky strongly criticizes Anglo-analytic philosophers for proudly taking basically all other insights from biology and neuroscience into account, while ignoring the science on agency. The Chinese philosophical approach, in contrast, allows us to appreciate precisely these factors and expound on their moral significance.[27]

Since the Enlightenment morality and ethics in Western thought has largely been dominated by two approaches: those influenced by deontological thinking, and those that concentrate on various types of utilitarian calculations. Both are fundamentally grounded in a hyper-atomistic and rationally charged view of the individual and the world—along the lines described by Zuboff above. Much of contemporary moral and ethical discourse is a continuation of these themes, and has come to be dominated by perspectives on the person, emotions, and agency as outlined earlier.[28] That are persons meaningfully composed of and with their relations, that emotions are important, the environment influential, and agency is not an abstract rational power, are all hugely downplayed in the moral and ethical discussions of Anglo-analytic philosophers. As will be demonstrated in the

26 Byung-Chul Han elaborates this in ways that allow us to appreciate Zuboff's point in a broader context. Han writes: "Big data and artificial intelligence enable the information regime to influence our behavior at a level that lies below the threshold of consciousness. The information regime takes hold of those pre-reflexive, instinctual, emotive layers of behaviour that precede conscious action" (Han 2022, 10).

27 There are many Western philosophers who also conceive of the person and agency along similar lines, with Nietzsche as one such example.

28 For example, Nick Bostrom, a key figure in AI theory and AI ethics, is an extreme consequentialist (see Bostrom 2014).

following section, AI researchers have taken their cues from these philosophers, and developed algorithms accordingly. Perhaps most damningly, the attitudes taken towards AI are equally based on assumptions about humans and the world as nothing more than highly complex mechanical processes[29] that can, and will, be meaningfully expressed and steered by powerful computers with highly "intelligent" machine-learning algorithms.

Moral and ethical views coming out of early Chinese thought reject this orientation in thinking about human interactions. Confucianism tells people to start with living up to their roles and cultivate themselves therein. Beginning with immediate family roles, the person is supposed to gradually learn how to act appropriately in various social contexts. Doing so always means thinking about context, environments, relationships, roles, others, and of course ritual and tradition. Reason is not given priority over emotions, nor are the two meaningfully separated. While there are specific virtues to be cultivated, they are best thought of as achieved within interpersonal relationships. No one can be "filial" without parents—and how this plays out is ultimately dependent on the particular individuals involved. It is the parents and child who together in their relationship accomplish instances of filial piety. The "cultivation" a person experiences is of their reason, emotions, and agency as tending towards filial interactions. However, this requires constant vigilance. No one ever "became filial"—they can only be said to excel at manifesting good relations (or "filial" ones) with others. And since transformation is constant, and all aspects of the environment have influence, complexity is constant and unresolvable. Confucian ethics is best classified as ways of reflecting on interpersonal relationships. Unlike deontological thinking, utilitarian calculations, and the mechanistic treatments of Anglo-analytic philosophy, early Chinese thought neither offers, supposes, or even desires final *solutions*.[30] In this way it is quite unfit for discussion in contemporary academia. Some of the most famous phrases regarding ethics, morality and the complexity they entail include: "Humans broaden the way, the way does not broaden humans" (*Analects* 15.29), "that was one time and this is another" (*Mencius* 2B:22), and "there is nothing I must do and nothing I must not do" (*Analects*

29 AI theorists and ethicists often translate human abilities into computer terms. For example: "Biological neurons operate at a peak speed of about 200 Hz, a full seven orders of magnitude slower than a modern microprocessor (~ 2 GHz)" (Bostrom 2014, 53).

30 Commenting on an early draft of this paper Dimitra Amarantidou writes: "It's not that we don't want solutions in Confucianism. But we don't want one solution or one set of reiteratable solutions. It's not a single-solution theory, right? We want a proliferation of context-dependent solutions. And this is why it is unfit for current discussions. Scientific thinking can only progress on a single-solution basis. Because it can only focus on one, as narrow as possible, problem at a time. Kongzi [Confucius] is dealing with many questions at a time when he talks with people. And he expects them to think in this multi-polar way, too."

18.8). In the realm of contemporary academia, but especially from the perspective of Anglo-analytic philosophy, these are incredibly unsatisfying statements. However, they are not supposed to convince through argumentation—or, their "argument" can only become manifest in considering these ideas when living and experiencing their validity. That is how Chinese philosophy has always operated, and it makes it uniquely applicable to dealing with the "spontaneous experience of reality, in all its contradiction and complexity".

Math and Meaning

"The idea that society can be made more consistent, more accurate, and more fair by replacing idiosyncratic human judgment with numerical models is hardly a new one"—so opens Brian Christian's chapter on "fairness" in his book *The Alignment Problem* (2020). The "alignment problem" is a widely discussed issue in AI research. It refers to the problems involved in building AI systems that function in ways which are aligned with the expectations of their developers or users. Often this is discussed in terms of values, what Stuart Russell calls the "value alignment problem": the problem of trying to align AI values with human values. Many AI researchers turn to formalized metrics, statistical specifications, and other numerical models to achieve this, as Christian suggests. In a generalized sense, this is exactly the language of Anglo-analytic philosophy, and we might gloss it as a mathematical approach as well.

To a limited extent improving mathematical models can improve society, but mainly in areas where math is foundational. For example, the terrible bias in facial recognition technology turned out to be a problem of data. The data sets used to train this technology were largely comprised of white male faces, making predictions of female and especially non-white female faces extremely poor.[31] Here making AI align better with human values, or more "humanistic", can be seen as a math problem. Better data sets will mean more accurate predictions. However, the humanistic consequences of predicting that a human face is actually the face of a "gorilla" is not easily translated into 0's and 1's. While some, including Christian, have suggested that we can use math to make AI "understand" the qualitative difference between misidentifying a cat as a dog and a human as a gorilla. The problem is far from mathematical. Environments, interactions, and contexts all play a role in complex and unpredictable ways. Formalization restricts or even kills values. We need to be extremely careful when values and math intersect, be it in AI or in humans.

31 Research related to "WEIRD" or "Western educated industrialized rich and democratic" has shown that many scientific studies are skewed towards a narrow group of people.

In his book *Parentonomics* (2010), Joshua Gans refers to his role as a father as "one big economic management problem", and is quite serious about using incentives and rewards to raise his children. Food, he says, worked as a great incentive for his daughter, and so he used it in various ways to reward her for desired behaviour. For instance, he rewarded his daughter with a piece of candy every time she helped her younger brother go to the bathroom. An incentive to help with the potty training process. The daughter reacted, quite unsurprisingly given the environment, with the same mechanistic thinking her father applied, and promptly began to force her brother to drink as much water as possible. More trips to the bathroom equalled more candy. The same type of exploitation of unintended loopholes famously happens in AI research all the time. There are countless examples. One is of Astro Teller and David Andre's soccer program. In an effort to make better AI players Teller and Andre programmed a reward—far smaller than that for scoring—for taking possession of the ball. "To their astonishment, they found their program 'vibrating' next to the ball, racking up these points, and doing little else" (Christian 2020, 167). It could easily "add up" more value with possession than it could score.

Commenting on how we might reflect on the "incentive failures" in the bathroom and soccer examples given above, Brian Christian finds hope in recent AI developments: "Cognitive sciences and economists are turning to computer sciences to develop incentive structures that do not distort behavior" (Christian 2022, 25:10). For example, to correct the older sister from forcing her younger brother to drink cups of water, Christian suggests she be scolded "in precisely equal measure … [for instance taking a candy away when she forces her brother to drink water] so that the net gain of further repetitions is zero" (Christian 2020, 170). This works with AI, but that does not mean it should be applied to humans—even if some desired results do occur.

Before dealing with the development of appropriate corrections we need to recognize that that which allows us to immediately realize the "incentive failures" is not computer science at all. That they are instead examples of incentive structures gone wrong is immediately evident to everyone. No math is needed to achieve this—in fact, no math can bring about this conclusion at all. It is exactly idiosyncratic human judgment that allows us to agree. No one has to explain or prove why and how forcing a small boy to drink water does not constitute "being a good sister" or robots vibrating soccer balls is not "gaining possession of the ball", and far less "playing soccer". And no one really could, at least not in a way that would elicit universal agreement.

If we did try to translate what we immediately recognize as a blunder on the part of the big sister or the robots into the mechanistic language we would quickly find ourselves in a difficult position. Asking the big sister to replace one mechanistic model (helping her brother go to the bathroom equals candy) with another model comprehensive enough to ensure no more loopholes would be extremely time consuming, if even possible. How could one realistically account for every foreseeable situation, and get a young girl to understand? In fact, it is no more possible than it would be meaningful or practical.[32] The world is complex, and straightforward mechanistic rules only help us in the direst situations, or with issues like traffic or tax laws. In other areas excessive formalization often fails miserably.

There is a critical difference between asking children to say "thank you" or "sorry" in the Chinese realm (or the realm of most human parents) and providing "incentive structures" or "parentonomics". In actual practice the two may overlap, but how they approach the problem and how they seek to deal with it are critically different. If humans are atomistic, reason-based, with an isolated power of autonomy, and ethics can be broken down into codes, then how we treat people, what we expect from them, and how we view their development follows certain lines. Based on these assumptions it makes sense to look to computer science and the most efficacious algorithms in AI for guidance. We might try, as Christian suggests, to "zero out incentives". In doing so, however, we need to recognize that we not only treat people like mechanistic mathematical entities (or robots) but we expect them to be so as well.

The downside is clear. When this type of thinking dominates people can and do tend to look for loopholes, they satisfy the measure in a hollow or even counterproductive fashion[33] or "follow the letter of the law and not the spirit". As Stuart Russell argues, if we try to reduce everything to mathematical codes, we should remember that for thousands of years tax law has been trying to close off every loophole, and for thousands of years people have been successfully evading taxes. Drinking too much water and vibrating soccer balls are relatively innocuous examples, but what Russell warns of is these models dominating critical social systems, and then being doubly reflected in AI. In other words, we not only develop AI along these lines, but our correctives are similarly conceived. We try to fix problems by, for example, "putting in values" or developing ever more complex

32 Something vague such as: "don't let your brother drink too much" or "don't encourage him to drink too much" could be tried. But both can obviously fail in certain situations, and "too much" is something exceedingly difficult to define with any precision. Perhaps we should seek to program her with a different algorithm such as "help your brother when he needs help". Any parent knows, however, that children can often help one another doing things they are not supposed to.

33 "Goodhart's law" and "the Cobra Effect".

models, we do not address the core issues (Russell 2019, 177–79). Humans finding loopholes can be bad enough, from forcing younger brothers to drink too much water, to tax evasion and war. Depending on what we allow AI to influence or control, the loopholes it may find could lead to financial markets crashing, corrupting elections (again), and nuclear disasters.

When the *Analects* states "ornamental aspects are like essential aspects and essential aspects are like ornamental aspects" (*Analects* 12.8) the "proof" of this argument is given in an example—one which draws not on reason or logic, but on experience and emotion-infused thinking. Again, it is all about the concrete and all about drawing on a broad range of particulars. The example given to "argue" the point is: "The hide of a tiger, when stripped of its hair, is like the hide of a dog or goat stripped of its hair". The ornamental aspect is the pattern made by fur. The "essential aspect" is the actual hide, the part that keeps one warm. In nearly all situations the hide of a tiger would be seen quite differently than that of a dog or goat. The former is likely far more expensive and could connote prestige, wealth, and the like. The only time we would think of them as the same, that is, the only time we would not consider their social evaluation and the meaning of their ornamentation, is in the direst of situations. In some remote area, lost or cut off from society.

From this perspective teaching a child to say "thank you" or "sorry" is not about "incentives". The goal is *not* to develop a comprehensive and robust enough structure of rewards and punishments in the hope of guaranteeing that the child's "idiosyncratic human judgment" can be replaced by a "more consistent, more accurate, and more fair" numerical model, nor does "zeroing out" incentives come into play. Agreeing with Kissinger, Schmidt, and Huttenlocher, Chinese philosophy sees the humans and the world as spontaneous, and full of contradiction and complexity. We teach a child to say "thank you" or "sorry" in hopes of developing their persons—orienting the growth of their emotions, reason, and agency. We try to help them figure out what environmental markers are most important, and give them models for reflection when they encounter new situations in the future. Truly, every situation is unique to some extent, so we need to think in ways that allow us to be continually critically reflective. The cultivation of a person is about expanding. Mathematical models are necessarily narrow, and this is a far more precarious strategy for everything from parenting children to programming AI.

Indeed, when Christian suggests that numerical models can aid in social consistency, accuracy, and fairness, he is not wrong. Cathy O'Neill also believes that math can be used to improve society along these lines. Again, it is not mathematical

models themselves that are the issue. The issue is how they are used, and whether we reflect with them to make the world smaller and narrower, or to broaden our perspective and take particulars into account. In other words, better mechanics, better math, and more powerful machines can iron out some problems, but only to a limited scope, and to a limited degree.

The *Zhuangzi* 莊子 (*Book of Master Zhuang*) relates the story of "chaos"—which we can understand as an appreciation of the spontaneity, contradiction, and complexity of the world. Chaos was a "nice" (*shan* 善) host to the rulers of the northern and southern oceans who would often meet in Chaos' realm. One day the rulers decided that Chaos was being "virtuous" (*de* 德), and according to their formalized and mechanistic thinking of virtue, they were supposed to "repay" (*bao* 報) Chaos' "virtue". "Since every person has seven orifices and Chaos has none, we should give him some" they reasoned. After drilling the seventh hole, Chaos died (*Zhuangzi* 7.7).

This story lends itself to a wide variety of interpretations. But one of the most basic messages we can glean from it is that we destroy spontaneous "niceness" by putting it into formalized and mechanistic systems. Confucianism too, even while promoting rituals, guards against overly formalistic and mechanistic tendencies. For example, the *Mencius* says ideally one should act *from* humaneness and duty, one should not act *according to* humaneness and duty (*Mencius* 4B: 47). In other words, one is not supposed to behave mechanically according to formalistic rules, rather one should actually develop themselves into a good person. This is exactly what we are teaching children when we force them to say "thank you" or "sorry". We are trying to teach them to go beyond the type of thinking that leads children to force brothers to drink (or robots vibrate next to soccer balls). We want children to appreciate the *meaning* of "thank you" and "sorry", and they do not do so by learning math problems.

Conclusion

Describing his book *Ethical Machines* (2022), which claims to be a "Guide to Totally Unbiased, Transparent and Respectful Machines", Reid Blackman writes: "don't worry—the book's purpose is to get work done, not to ponder deep and existential questions about ethics and technology. [My] clear and accessible writing helps make a complex and often misunderstood concept like ethics easy to grasp" (Blackman 2022, 225). Indeed, this is the dominant approach. Michael Kearns and Aaron Roth, authors of *The Ethical Algorithm: The Science of Socially Aware Algorithm Design* (2019a), also take issue with philosophical discussions of

terms such as "privacy", "fairness", and "morality". They argue that we must explain words like "privacy" "not in the way a lawyer or philosopher might describe them, but in so precise a manner that they can be 'explained' to a machine" (Kearns and Roth 2019a, 18).

"Privacy" is key for Kearns and Roth because they think algorithms can already be programmed in a way that ensures privacy—and they can mathematically prove it. Not just any privacy, however, what they discuss is "differential privacy" which is clear enough to be "explained to machines". They define such privacy as "a mathematical formalization of the [idea] that we should be comparing what someone might learn from an analysis if any particular person's data was included in the dataset with what someone might learn if it was not" (Kearns and Roth 2019a, 36).[34]

They further explain:

> The definition of differential privacy is, it's a property of an algorithm, first of all, not about a particular data set. An algorithm is or is not differentially private. And differential privacy is generally achieved by adding noise to computations. […] You add noise that obscures the contribution of any particular data in the analysis while preserving on statistics. (Kearns and Roth 2019b, 18:15)

Because it works in practice Kearns thinks this is a "great definition"[35] despite even its broad generalizations—for example, he says "the definition of harm can be anything you want it to be" (Kearns and Roth 2019b, 8:30). Of course anyone who "ponders deep and existential questions", any lawyer or philosopher, would have a field day with Kearns' "harm". Nevertheless, he is not entirely wrong.

In a very real sense Kearns and Roth, along with Blackman and Christian are all correct. While from a humanistic perspective we might want ethics to be about meaning and linked to human experiences of the world, there is not really any way we can translate this into code. Even if loopholes were not the issue, the very meaning of *meaning* we have outlined above is precisely opposed to numerical

34 Wikipedia defines differential privacy as "a system for publicly sharing information about a dataset by describing the patterns of groups within the dataset while withholding information about individuals in the dataset. The idea behind differential privacy is that if the effect of making an arbitrary single substitution in the database is small enough, the query result cannot be used to infer much about any single individual, and therefore provides privacy," (https://en.wikipedia.org/wiki/Differential_privacy (accessed November 21, 2022)).

35 Kearn also states: "Sometimes the very exercise of having to think so precisely about the definitions of these social norms is itself greatly beneficial. It will not only reveal trade-offs that you were not aware of but it will also reveal flaws in your intuitions about these ideas if you just talk about them at the level of moral philosophers" (Kearns and Roth 2019b, 8:00).

models. Or, put another way, humanistic meaning itself is not translatable into algorithms. Kearns and Roth and Christian also appreciate the limits of their approach. Christian notes that models of the world are always imperfect—even if he assumes more accurate ones could be made with the right math, data, and computing power. Kearns and Roth also note that "the science can only take you so far".[36] But again, they all do take science to be an extremely important guide for human thinking and practice. In other words, they all recognize a certain paradox: AI challenges us with new questions of meaning and of human experience, but in exactly ways whereby AI itself can never be expected to make meaningful predictions. We thus find ourselves in a place where AI helps us make sense of the world, while at the same time making it more opaque and clouding our understanding.

Fortunately, some of the shifts in methodology and changes to static and atomistic assumptions about concepts such as people, emotions, agency, and ethics, and even our understandings of meaning and value, can be reflected in the way we program AI. It would require significant adjustments not only in the math involved, but *how* that math is used. This is possible, and some computer scientists are already building models which incorporate these perspectives.

Stuart Russell has been working on "inverse reinforcement learning" for over two decades. His most recent book contains a particular humanistic orientation, found already in the title *Human Compatible* (2019). Here he describes shifting AI programming from being centred on actions to states. This can be one way to understand the different thinking behind the examples of forcing brothers to drink water and children to say "thank you". The older sister thought in pure action-based mechanics, while saying "thank you" hopes to foster a certain state—both in the immediate environment and the individual. Russell describes three principles which are "intended primarily as a guide to AI researchers and developers in thinking about how to create beneficial AI systems; they are *not* intended as explicit laws for the AI system to follow" (Russell 2019, 172). (Here he already calls for moving away from mechanistic and formalized thinking, and is thus far closer to Chinese thought than to certain orientations built off Anglo-analytic philosophy.) The principles of Russell's

36 "The science can only take you so far. It can elucidate where trade-offs are [when balancing, for example, accuracy and bias in AI] but it cannot tell you where on the trade-off curve you want to live—as a society and in particular applications. And there are not going to be universal answers. We will want to prioritize fairness or privacy more in certain applications, we will want to prioritize accuracy in other applications. But there is no avoiding that we have to make hard decisions. What the science can do is help us make those decisions with our eyes open." (Kearns and Roth 2019b, 30:50)

beneficial AI systems are:

1. The machine's only objective is to maximize the realization of human preferences.
2. The machine is initially uncertain about what those preferences are;
3. The ultimate source of information about human preferences is human behavior. (Russell 2019, 173)

Russell then goes on to note that preferences will be different for different groups, for different individuals, in different situations, and change over time. Many complex issues, such as whose preferences should count, how we should weigh preferences on a social level, or there being some preferences that should not count, and other deep and existential questions, will have to be discussed. And those discussions will never end. It is an approach which necessarily does not provide *solutions* to ethical issues, but a broad and encompassing *approach*, one which takes into account the "spontaneous experience of reality, in all its contradiction and complexity". If Chinese philosophy has contributions to make in reflections about the use of AI, it is along the same lines as Russell's "inverse reinforcement learning".[37]

Indeed, while there are certainly good resources for injecting more humanism into AI in Western philosophy, and perhaps developing a new philosophy, as Kissinger et al. suggest, would be best suited for the unprecedented problems we face. In any case, the approach we find in Chinese philosophy is certainly a valuable resource for thinking about the humanistic problems with AI. In short, in collaborative approaches to AI research references to philosophy should expand beyond the Anglo-analytic tradition, and Chinese philosophy is a great counter to the prevailing tendencies, as outline in this paper.

Acknowledgement

This project was funded by 华东师范大学海外发文项2019ECNU-HWFW010.

37 It is worth noting again that the contributions to AI research that Chinese philosophical approaches stand to make do not wholly replace those made by Anglo-analytic philosophy, and both should be used in conjunction with one another.

References

Blackman, Reid. 2022. *Ethical Machines*. Cambridge, MA: Harvard University Press.

Bostrom, Nick. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Bryne, Sarah. 2020. "Predict and Surveil: Data, Discretion, and the Future of Policing." YouTube, accessed November 7, 2022, https://www.youtube.com/watch?v=Jo-3vRTPTDw.

Buolamwini, Joy. 2019. "AI, Ain't I a Women?" YouTube, accessed June 29, 2022, https://www.youtube.com/watch?v=HZxV9w2o0FM.

Buolamwini, Joy, and Timnit Gebru. 2018. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, PMLR 81: 1–15. https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf.

Christian, Brian. 2020. *The Alignment Problem: Machine Learning and Human Values*. New York: W.W. Norton and Company.

———. 2022. "The Alignment Problem: Machine Learning and Human Values with Brian Christian." YouTube, accessed November 23, 2022, https://m.youtube.com/watch?v=z6atNBhItBs.

Christian, Brian, and Tom Griffiths. 2016. *Algorithms to Live By: The Computer Science of Human Decisions*. New York: Henry Holt and Co.

Fridman, Lex. 2022. "Rana el Kaliouby: Emotion AI, Social Robots, and Self-Driving Cars | Lex Fridman Podcast #322." YouTube, accessed November 11, 2022, https://www.youtube.com/watch?v=36_rM7wpN5A.

Gans, Joshua. 2010. *Parentonomics: An Economist Dad Looks at Parenting*. Cambridge, MA: MIT Press.

Gebru, Timnit. 2020. "Race and Gender." In *The Oxford Handbook of Ethics of AI*, edited by Markus D. Dubber, Frank Pasquale, and Sunit Das, 251–69. Oxford: Oxford University Press.

Haidt, Johnathan. 2003. "The Moral Emotions." In *Handbook of Affective Sciences*, edited by R. J. Davidson, K. R. Scherer, and H. H. Goldsmith, 852–70. Oxford: Oxford University Press.

Han, Byung-Chul. 2022. *Infocracy: Digitization and the Crisis of Democracy*. Cambridge, MA: MIT Press.

Huberman, Andrew. 2022. "How to Maximize Dopamine & Motivation—Andrew Huberman." YouTube, accessed November 17, 2022, https://www.youtube.com/watch?v=ha1ZbJIW1f8.

Kearns Michael, and Aaron Roth. 2019a. *The Ethical Algorithm: The Science of Socially Aware Algorithm Design*. Oxford: Oxford University Press.

———. 2019b. "The Ethical Algorithm | Michael Kearns & Aaron Roth Talks at Google." YouTube, accessed November 13, 2022. https://www.youtube.com/watch?v=tmC9JdKc3sA.

Kissinger, Henry, Eric Schmidt, and Daniel Huttenlocher. 2021. *The Age of AI*. New York: Little, Brown, and Company.

Noble, Safiya Umoja. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: NYU Press.

———. s.d. "A Revealing Look at How Negative Biases Against Women of Color are Embedded in Search Engine Results and Algorithms." *Amazon*. Accessed November 17th, 2022. https://www.amazon.com/Algorithms-Oppression-Search-Engines-Reinforce/dp/1479837245.

Nussbaum, Martha. 2010. *Not for Profit*. Princeton, NJ: Princeton University Press.

O'Neill, Cathy. 2016. *Weapons of Math Destruction*: *How Big Data Increases Inequality and Threatens Democracy*. New York: Crown Publishers.

Roberts, Huw, Josh Cowls, Jessica Morley, Mariarosaria Taddeo, Vincent Wang, and Luciano Floridi. 2021. "The Chinese Approach to Artificial Intelligence: An Analysis of Policy, Ethics, and Regulation." *AI & Society* 36: 59–77. https://doi.org/10.1007/s00146-020-00992-2.

Rošker, Jana S. 2014. "Ji Kang's Essay 'Music has in it Neither Grief nor Joy' (聲無哀樂論) and the Structure (理) of Perception." *Philosophy East and West* 64 (1): 109–122. http://www.jstor.org/stable/43285882.

Russell, Stuart. 2019. *Human Compatible: Artificial Intelligence and the Problem of Control*. New York: Penguin Books.

Sapolsky, Robert. 2021. "Dr. Robert Sapolsky: Science of Stress, Testosterone & Free Will | Huberman Lab Podcast #35." YouTube accessed November 27, 2022, https://www.youtube.com/watch?v=DtmwtjOoSYU.

Simanowski, Roberto. 2018. *The Death Algorithm and Other Digital Dilemmas*. Cambridge, MA: MIT Press.

Turow Joseph. 2017. *The Aisles Have Eyes*. New Haven, CT: Yale University Press.

Wu, Tim. 2016. *The Attention Merchants*. New York: Vintage Books.

Zuboff, Shoshana. 2018. *Surveillance Capitalism*. New York: Public Affairs.