Digital and Computational Aspects of Human and AI Ethical Frameworks*

Hector Zenil

Algorithmic Dynamics Lab

Unit of Computational Medicine, Centre for Molecular Medicine, Karolinska Institute, Stockholm, Sweden & Algorithmic Nature Group, LABORES for the Natural and Digital Sciences, Paris, France

Abstract

I explore previous attempts, including recent ones, to introduce aspects of digital information and computation into the discussion of ethical frameworks. I study some limitations and advantages of previous attempts to produce guiding principles at different scales. In particular, I survey and discuss questions and approaches based on, or related to, simulation, information theory, integrated information, computer simulation, intractability, algorithmic complexity, and measures of computational organisation and sophistication. I discuss and propose a set of desirable features of ethical frameworks that may be considered well-grounded, both in theoretical and methodological terms. I will show that while global ethical frameworks that are uncomputable are desirable because they provide non-teleological directions with open-ended meaning, constrained versions should be able to provide guidelines at more local and immediate time scales. In connection to the ethics of artificial intelligence, one point that must be underscored in relation to computational approaches is that (General) AI should only share and embrace an ethical framework that we humans are willing to adopt, preempting the need for possibly flawed distinctions between human and machine actions, especially in connection to concerns of a much more fundamental nature than the classical issues raised by AI development such as job displacement and legal liability. I think that such framework is possible by following a general and universal (in the sense of computation) framework from first computational principles.

Keywords: philosophy of information, organised complexity, Kolmogorov complexity, logical depth, ethics of information, computational ethics, infoethics, machine ethics, computational complexity.

---

*hector.zenil [at] algorithmicnaturelab [dot] org

1   Ethical Questions

Humans face ethical and moral challenges every day, in the form of conflicts over what is best in the short term versus what may be best in the long term, or what may be best for oneseld or one's group versus what may be best for others. For example, consider travelling: one may want to travel somewhere for leisure but refrain from doing so to limit pollution. Or smoking: one may decide to smoke for short-term satisfaction despite the likely long-term consequences. Some people choose not to care about pollution, others decide not to travel, some to smoke, others to refrain. We are constantly faced with such dilemmas.

One interesting and recurrent case in this manuscript will be antinatalism, an extreme position that establishes that avoiding human reproduction is the appropriate ethical choice when it comes to the question of procreation. The position is interesting because in a local scenario the logical reasoning seems flawless and can only be refuted if non-local arguments are brought to bear. There is also a matter of time scales of a binary nature, either someone exists to be taken into consideration when making a choice or they do not.

There are a number of ethical frameworks which philosophers have conceived, from Aristotle and his virtue-based ethics [2], Kant's deontological ethics [15, 16, 17] and Mill [20] and Bentham's teleological ethics [7]. Although formal interest in ethics can be found in the development of approaches such as deontic logic, ethics and logic have for the most part diverged, separating the humanities from science. Toby Walsh, a Professor of Artificial Intelligence, has expressed what appears to be a pragmatic observation, namely, that if humans cannot specify their ethical and moral codes with precision, we cannot expect to find it easy to do so for AI:

> For too long philosophers have come up with vague frameworks. Now that we have to write code to actually implement AI, it forces us to define a much more precise and specific moral code

Here I will advocate that for ethical frameworks to be relevant, translatable, and eventually readable by machines, human ethical frameworks need to be formalisable, and that computational approaches offer such transferability. My proposal is thus fundamentally different from those based on logic and probabilistic foundations such as in [1], and does not make an ontological commitment in any particular framework other than proposing that frameworks be evaluated and enriched by a set of concepts and features related to computability. Instead, I will explore some suggestions regarding computational ethics [21] and suggestions for introducing measures of complexity or sophistication into ethics [14, 6, 12].

A first question concerns what to consider a valid ethical question and a valid framework within which to place it or derive it from, a question about which a human being can make a conscious decision that will have consequences for others as well as for him/herself. For example, is the question of procreation a valid ethical question? It should be, because we are the only species on earth

able, thanks to biological evolution, to choose whether to reproduce or not, aided by such technologies as contraception. This means that an ethical framework such as antinatalism is a valid and legitimate one.

I will explore the direct and indirect implications of developments attempting to introduce digital information and complexity indices into the ethical discourse.

There are also questions that should not concern ethics, the ethical answers to such questions being meaningless. For example issues arising from anything that is accidental, such as, say, lightning striking trees, because in principle (unpredictability) and in practice, no human action can prevent such an occurrence (protect all trees or prevent storms).

There are also no absolutely (un)ethical questions. For example, it could be conceivable that within some framework, and perhaps (but not necessarily) on the basis of hard evidence, it is decided that plants do have some sort of internal experience and can feel pain, and that humanity finds it within the realm of the possible to protect all trees from lightning. Then such a question/concern may become ethical. Within some ethical/moral frameworks cannibalism or incest may not be unethical but in others they are. Likewise the way in which animals may be killed (e.g. halal versus non-halal). If no absolute ethical/moral framework can exist, the value of and the answer to an ethical question is relative to the framework of reference.

1.1   Not all Frameworks are Equal

To begin with, no universal or objective ethical framework is possible. For example, had homo sapiens as a species not evolved to the degree that we have, human reproduction could not be said to be good or bad for the world or the universe. We have therefore to start acknowledging that all ethical frameworks are, and only make sense, if they are human-centred and are designed to deal with the challenges that already existing humans face or will face. A recent study clearly shows how not even principles that some populations may believe to be general are actually so. In a study related to moral and ethical dilemmas involving car crashes [3], Eastern countries were found to dramatically favour sparing the elderly and lawful pedestrians, while Western countries favour the young and have adopted some utilitarian preferences, leading them to prioritise sparing the greatest number of people above all else, followed by sparing people of higher social status and fitness. Thus Eastern countries dramatically diverge from Western and Southern countries, showing little concern for numbers saved, social status or looks. Southern countries, in contrast, do not favour lawful pedestrians over others and care mostly about sparing people with higher status, the young, the female and the fit, and show a much higher preference for sparing animals over humans than the other two groups. This clearly shows that different populations may embrace different ethical values and principles.

While any principle can be placed at the centre of an ethical framework there cannot be a way to falsify an ethical framework on the basis of any objective universal measure, because that would necessarily imply evaluating a

framework from the standpoint of an alternative framework, which leads to an immediate contradiction (the evaluating framework would be more primitive or fundamental than the purportedly universal evaluated framework). However, not all ethical frameworks are equal or should be equal.

An ethical framework can, for instance, be tested for internal consistency at different levels and time scales to see whether following its principles leads to contradictions. In this sense it is not, or should not be, that different from formal logic, in which consequences have to be compatible with the purposes of a given framework. A framework that often tells you that you have to follow certain principles because you cannot see the greater plan is the kind of framework I would propose be relegated in favour of less obfuscatory frameworks, for the sake of consistency at various scales and levels.

We should thus not equate plant signalling with a sentient being's pain and hence justify assigning equal value to ethical frameworks based on notions of plant 'pain' and those dealing with animal or human pain. We are sure that the latter is a consequence of an internal conscious experience, while the consensus is that plants have no such experience or anything indicating that they do. But, more importantly, the ethical question here concerning plant suffering is that plants do not have decision-making power with regards to their own pain (not to the extent of being able to name it and decide on measures to alleviate it). This means that the question of plant suffering may not be equally legitimate.

In contrast, humans have the ultimate capacity to stop all possible pain, even the ineluctable pain of existence, which makes antinatalism, that is, the decision to stop replicating and thus the decision not to continue existing as a species, a legitimate framework, whether or not it may be regarded as an extreme stance from the vantage point of some other ethical/moral framework.

We will thus agree that any principle can be arbitrarily placed at the centre of an ethical framework but that not all principles so placed are, or should be, equally valid.

One advantage of arbitrary (random) ethical frameworks, based, for example, religion, is that they do provide clear rules. However, whether they are sound and self-consistent is another matter, as we have seen how their doctrines can be shaped to fit different, often contradictory, purposes (e.g. killing in the name of god). So many of these are inconsistent and hard to follow, even making it difficult for bona-fide adherents of the same religion to understand each other.

Here, I will suggest that aspects of computability could help us decide what should be taken as desirable features and properties of an ethical framework, and also dictate computable strategies, even if these ethical framework may lead to uncomputability– which should be embraced and understood. As I will suggest, non-decidability and uncomputability avoid teleological, goal-oriented frameworks, allowing for a healthy number of different possible directions that can be adopted to attain more local and personal ends.

2   A Case Study: Antinatalism

An extreme case of an ethical dilemma will help us analyse some aspects of proposals for human and AI ethics of particular relevance to Artificial General Intelligence (AGI), which has been alleged to pose a threat to human existence [9]. The case concerns procreation but not suffering. What is in question is procreation, it being generally agreed hat suffering and pain are wrong. An ethical framework that says that pain or suffering is good or not relevant will have to be immediately challenged as non-relevant to humans.

Well-being and happiness can be regarded as a sum, as is done in traditional consequentialist, utilitarian ethical frameworks. Well-being or happiness versus pain are so opposed that it is safe to regard their relationship as a zero-sum game, where if H is happiness and P is pain (of any form, such as physical or mental suffering), then $H - P = 0$. Interestingly, the number of players changes according to the perspective. From the point of view of the unborn, there is only one player in the zero-sum game. I claim that any other framework in which the only relevant point of view is not only that of the unborn is selfish because the unborn cannot make the choice, hence by definition selfish because any choice would be exclusive of other players. Once adopting selfishness the framework becomes a non-zero game with more actors and greater complexity. In the altruistic case, the scenario is binary because there are only two options: either the newborn exists or it does not. From the perspective of the unborn/newborn, it may have H happiness and P pain/suffering with H and P being real numbers. However:

1. If a human being does not yet exist it does not make sense to ask if a non-existent person can be H happy, though one can know with certainty that by not existing $P = 0$, thus no suffering. However,

2. If the human being is born, then only $H = 100\%$ can justify the choice. Therefore, because $H = 100\%$ is unlikely or even impossible, then only $P = 0$ can be justified and the person should remain unborn in an altruistic world.

Option 1 is the non-existence option; it minimises pain and suffering. The question of happiness has no meaning because the person does not yet exist but once they exist then H and P have meaning, and P is certainly not 0 in their lifetime. The rules that apply to someone who does not yet exist do not apply to someone already forced to exist. In other words, noting can be bad or wrong for someone who does not yet exist but everything could be bad or wrong for someone who does. Hence bringing anyone into being who cannot participate in such a decision cannot but be a selfish act within the isolated sphere of the parents-newborn relationship (which, of course, only comes into being after the parent makes the decision to have a child). But if external arguments are brought to bear, justifications can be found, though no external justification can be an internal unselfish argument at lower levels (see Fig. 1).

So the only way to reduce pain and suffering with absolute certainty and remain altruistic or not selfish is by choosing an antinatalist position, because
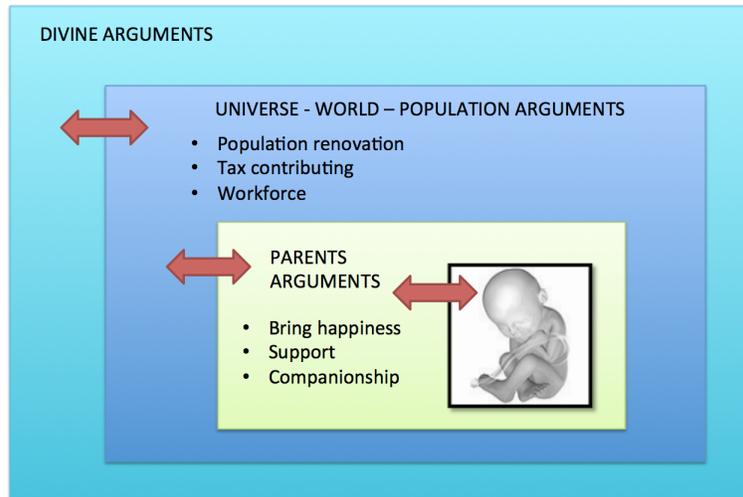
Figure 1: There seem to be apparent absolute selfish/unethical actions illustrated by, e.g., a parent-new born closed system, where one individual takes a long-term and far-reaching decision on behalf of another one by forcing it to live. Positive values amounting to a real number P between 0 and 100 cannot be assigned to an unborn because they do not yet exist. A $P > 0$ value can only have meaning after birth, but comes with a negative value N regardless of the specific values of P and N. The action can therefore be only justified by inserting it into a larger ethical/moral framework breaking the bubble where possible non-selfish reasons can be sought and advanced. Draw circles around the figure with even more weirder ethical frameworks, from population renovation, people who pay older people pensions, to divine justifications. Paint them according to their computable/recursive value, favour the inner circles, the more exotic the more likely uncomputable and algorithmically random (a possible recursive description of god would likely appear random because it can only be simple at the level of the god itself, under a divine explanation but not at a mundane level). The bubble is computable or semi-computable, at a local scale, all other things being equal. We should favour decidability but in the longer term we may embrace semi-computable theories (such as LD).

existing will always entail both happiness and pain, and parents are responsible for both when they could have avoided pain altogether, and the question of happiness makes no sense prior to the existence of the unborn, meaning that the decision to procreate was taken in a selfish manner.

In terms of Benatar [4], the argument follows that coming into existence generates both good and bad experiences, pain and pleasure, whereas not coming into existence entails neither pain nor pleasure. The absence of pain is good, the absence of pleasure is not bad. Therefore, the ethical choice is weighted in

in favour of non-procreation.

No amount of happiness can counterbalance the minimum amount of pain and suffering when the option is non-existence.

From a population point of view one can find ethical justifications, but not for the individual who is forced to live. One has to divide the question between the individual and the population as well as consider time, as it is not the same to force someone to exist and then ask them to kill themselves versus simply deciding not to bring them into being the first place.

Some may be willing to think of an artificial replacement for human existence but may find it shocking to consider that a possible good outcome can also be the disappearance of the human race, thanks to local decisions of humans not to procreate (as it is happening, see Fig. 2).

The question of procreation has an interesting logical side, and arguments in favour or against make clear how one has to devise an ethical framework able to answer or adapt to different scales. In this case, and by definition, deciding for someone else cannot be seen as universally ethical because forcing someone to exist is, by definition, a selfish act between parent and child. Outside the parent and child bubble, one can find all sorts of secondary justifications to make the decision appear ethical, and it may be so, but it cannot be between parent and child. The exercise also shows how ethics cannot be universal in all senses, being relative in time and hierarchy, as it is not the same to take a decision not to procreate as to later ask the baby or the parent to kill that very human being, nor is it the same to locally behave non-selfishly and globally face the consequences of local ethical rules. In other words, ethical frameworks can only be consistent if they are multiscale and hierarchical, and this should be a strong guiding principle, as suggested by Ruvinsky [21]. And I claim this is also a necessary condition for a practical AI ethical framework.

A general counter-argument to antinatalism is its possiblly self-defeating or inconsistent nature. If everybody adopts it, the consequence would be that the human race would quickly become extinct. Is that really what antinatalism wants or would embrace? This mode of reasoning is actually the mechanics of Kant's categorical imperative: if what you profess and do cannot be universalised, then it should not be followed. Perhaps what Kant meant was that local rules have to be consistent with global rules, but this shows that there is a question of scales and that hierarchies matter. For antinatalism, what goal/consequence would be more important? To preserve humans, for whatever reason, or to minimise pain? Lying is a comparable example: if everybody starts to lie, we would live in a horrible world. Not procreating is not generalisable.

3   Digital Information and Computation

3.1   Aspects of computational irreducibility

Another important point to consider in the discussion of computable ethics or ethical frameworks based totally, or partially, on computational (or constructible
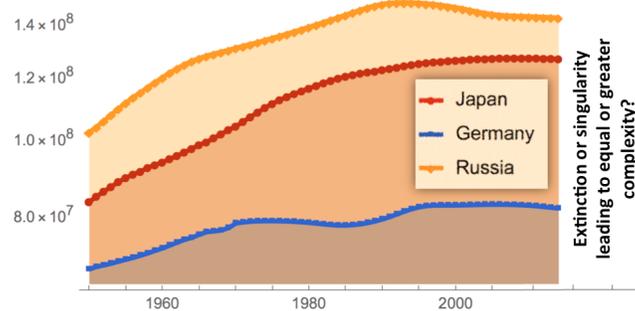
Figure 2: Diminishing population trends (Y axis). In developed countries (discounting immigration) over time. The same trend occurs among people with a greater number of years of education, even in developing countries, suggesting that the more educated, the less they procreate. If equally complex robots do not replace these populations, humans will not only go extinct but won't be replaced by anything more or equally sophisticated, such as machines, effectively leading to the realisation of the logical endgames of both antinatalism and digital antinatalism. The trend may also have more far-reaching implications as a possible answer to the Fermi paradox.

logical) approaches is the many forms of computational irreducibility that they may face. The strongest type of computational irreducibility is that imposed by the Halting problem, from which many others are derived, but weaker forms of irreducibility [24], including intractability, fall into this category too. For example, in pursuing the idea of agent-based simulations one has to be resources-aware. More pragmatical relationships between resource-bounded simulation and ethical frameworks has barely been investigated [21] but many angles are still uncovered, especially the more fundamental connections in particular to theoretical concerns. Here, I propose an extension to the simulation-based approach to ethics in the form of the adoption, introduction, and study of other features such as a combination of computable and semi-computable indexes as guiding principles, so that any human or machine is able to follow the principle rationale step-by-step, amounting to what I think should be a form of Turing-universal empathy (Fig. 3). Notice that by simulation here I do not mean only computational but also mental simulation. There being no evidence that the mind is non-computable (for a discussion see [25]), and since we have evidence that it can perform Turing computation, agent-based simulation is pertinent to both machines and humans.

A recent experiment along the lines of computable simulations leading to computational ethics [21], mostly mental but also aided by a computer imaginary and interface, [3] yielded extremely interesting results that allowed a classification of subjects that people believed would be more dispensable than others when emulating witnessing a situation in which a self-driving car crashes

Turing-Universal Empathy Scheme
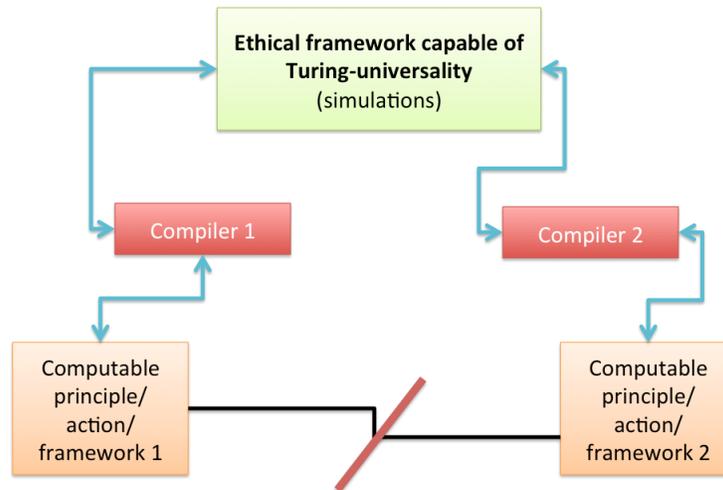
Figure 3: Universal empathy suggests that a Turing-universal ethical framework can mediate between other frameworks, actions and guiding principles, even when direct communication may not be possible. This is by way of compilers that are guaranteed to exist by virtue of computable systems. The compilers translate principles, actions or frameworks in terms of other frameworks. The ethical framework capable of Turing-universality is also capable of performing algorithmic simulations of all actions, guaranteeing that other systems can follow them step-by-step.

and kills passengers and/or pedestrians. Indeed, by simulating the outcomes of these car accidents, a popular vote ranked subjects according to their preference for saving different people and animals. A gender bias in favour of sparing females was significant. And between a dog, a criminal and a cat, people chose to save the dog. While this may look like a game, it is most relevant to current technology, where these preferences could be hard-coded in driver-less cars to decide who should live or who should die, though questions about liability would still remain. Furthermore is consensus the right criterion when it comes to evaluating ethical frameworks? We have seen how democracy in the course of human history has sometimes brought about undesirable outcomes, and that the miscarriage of democracy can be incredibly dangerous. Likewise wrong ethical frameworks even reached by consensus may also have undesirable consequences, even amounting to existential risks. Is deciding what is right and wrong a matter of popularity?

In the end, computational irreducibility imposes the limits by which any approach to ethics or anything else is bounded. It means, for example, that

we can neither perform all possible simulations nor derive all possible consequences of such simulations, nor can we fully carry out, foresee or shortcut most computations, and it is likely that final consequences can only be seen and witnessed as they actually unfold in real-time. Keeping these facts in mind, we can explore some measures for dealing with different kinds of irreducibilities, from time-space bounds to semi- and full uncomputability.

Beyond the introduction of logic, which I would regard as a form of computation when provided with constructible proving methods, there have been mostly 3 areas in which digital information and computation have been introduced in the last decade, from the pragmatical or theoretical point of view, from frameworks based on agent-based simulations that allow in silico testing [8], to ethical approaches heavily based on information theory [14] and a novel development based on complexity [12], itself based on previous suggestions [6]. This work adapts and builds on some of these previous ideas, except for information theory, which we would rather avoid because it is unsound, being highly dependent on the feature, language or probability distribution selected [**?**].

From measures of information allegedly able to quantify consciousness to recent proposals for ethics based on information theory and complexity, we have available to us possible resources to use in deriving global guiding principles from informational and computational first principles that may describe or explain interesting aspects of human activity.

3.2   Information-theoretic approach

The only approach that avoids the strongest form of irreducibility, that of semi- and uncomputability, is based on a weak but computable measure. This approach, based on classical information theory [14], proposes to place at the core of the ethical framework the purpose of minimising Entropy as a central guiding principle, in a sort of new age ecological approach to ethics. The idea is that agents, including humans, should 'avoid entropy' and oppose it with 'poiesis', that is, with construction. To be good agents in the infosphere, therefore, this approach suggests that we should be combating metaphysical entropy and promoting poiesis.

This approach leads to four ethical principles according to Floridi:

1. Entropy ought not to be caused in the infosphere,
2. Entropy ought to be prevented in the infosphere,
3. Entropy ought to be removed from the infosphere,
4. The flourishing of informational entities as well as of the whole infosphere ought to be promoted by preserving, cultivating, and enriching their well-being.

This position and approach, however, has many fundamental problems and is rooted in a fundamental lack of understanding of both thermodynamic entropy and Shannon entropy as a measure. Entropy is not a feature-independent measure and therefore before minimising entropy one has to make an arbitrary

choice as to the feature of interest in relation to which entropy is to be measured. Additionally, while minimising entropy would not lead to inaction, because by the 2nd law of thermodynamics it is impossible to avoid generation of heat from low entropy states in their progress toward higher entropy states in the environment, it may lead us to triviality. Indeed, entropy does not entail structure, and it can easily mean shallowness.

3.2.1   New challenges from integrated information

Along these lines, there is another interesting development connected to information theory in relation to dynamical systems. This is the concept of integrated information, suggested to be related to consciousness. Indeed, Integrated Information Theory, as introduced by Tononi et al. [23], not only proposes an apparent objective measure of consciousness but also poses new ethical questions.

For example, the measure ϕ related to Integrated Information, suggests that consciousness is a property of systems and not only of livings beings, and it also suggests that it is graded over continuous values, meaning that there is no such thing as a medical or scientific answer but rather an arbitrary cutoff value in attempting to decide whether a sentient being is conscious or not. It may, for example, open up a new debate related to abortion and the number of weeks after fertilisation when a fetus can be considered conscious, and it may even provide a numerical answer to the question of whether newborns attain the kind of consciousness that adults are endowed with and when they do so. The stages from fetus to developed child include several cognitive milestones, such as the awareness of their own body and of existence itself (which often causes pain when such existence is learnt to be finite), with body awareness beginning early and progressing gradually [13] until maturity. The measure ϕ is a continuous real-value number that goes from 0 to some as yet unknown number representing the typical (average) adult level of consciousness. The same would apply to questions related to passive or active euthanasia for cognitive conditions and neurodegenerative diseases. With ϕ not only indicating that consciousness is a dynamic state of the mind but also that it may be low enough to be below an abortion threshold, it could also supply the wherewithal to answer ethical questions about euthanasia.

Integrated information's ϕ is an interesting measure that may capture a necessary condition, but it is unlikely to quantify a sufficient condition (another clearly necessary condition is embodiment, i.e. a system must actually be embodied and capable of interacting with its environment, closing input-output loops). So I am not endorsing any suggestion that ϕ alone actually quantifies either consciousness or other possible factors that may be either intrinsic or extrinsic (the value of a person to a family, to society, etc).

Being a measure of systems rather than beings also implies that there may be artificial systems with some degree of consciousness, even if not completely related to ours. And there is no reason under this framework to exempt machines from certain degrees of consciousness, perhaps even equal to or even greater than

that of animals and the human being, which leads to all sorts of new ethical questions related to AI, and how to better approach ethics from computational perspectives, just as ϕ itself does.

3.3   An Approach based on Logical Depth

A more robust approach to an ethical framework, based on complexity rather than on classical information, was initially introduced by Bennett [6], based on his seminal concept of Logical Depth [5], a measure of 'sophistication' that quantifies irreducible computational work (see Appendix).

One criticism of algorithmic complexity is that it does not conform to our commonsense intuition of complexity. For example, when we say that something is complex we rarely mean that it is random, but actually that it is quite sophisticated, and neither trivial nor random. A measure of the complexity of a string can be arrived at by combining the notions of algorithmic information content and time. According to the concept of Logical Depth, the complexity of a string is best defined by the time that an unfolding process takes to reproduce the string from its shortest description. The longer the time it takes, the more complex. Hence complex objects are those which can be seen as "containing internal evidence of a nontrivial causal history".

Unlike algorithmic complexity, which assigns randomness its highest complexity values, logical depth assigns a low complexity or low depth to both random and trivial objects. It is thus more in keeping with our intuition of complex physical objects, because trivial and random objects are intuitively easy to produce, have no lengthy history, and unfold very quickly.

Logical Depth was originally suggested and advanced as the appropriate measure for evaluating the complexity of real-world objects such as living beings. Hence its alternative designation: physical complexity (used by Bennett himself). A persuasive case for the convenience of the concept of logical depth as a measure of organised complexity, over and against algorithmic complexity used by itself, is made by Bennett himself. Bennett's main motivation was actually to provide a reasonable means of measuring the physical complexity of real-world objects. Bennett provides a careful development of the notion of logical depth, taking into account near-shortest programs as well as the shortest one—hence the significance value—to arrive at a reasonably robust measure.

More recently, the proposal to base ethics on grounds of Logical Depth has been expanded [12] (I will denote logical depth hereafter by LD). The framework favours computational work. In some respects it may lead to similar consequences as its classical information-theoretic alternative, such as the preservation of species and biodiversity, but it is based on much more robust grounds, and has the advantage, for example, of being independent of language (to some extent) as well as feature-independent, thus being universal in a formal sense. But a shortcoming, that mau also be seen as its strength, is its uncomputable nature. On the one hand, it makes for open-ended meaning as it cannot be fully calculated and has no 'end', but at the same time it may lead to inaction. Moreover, it can lead to waste, for example, the energy (in the form of elec-

tricity) waste generated by the mining activity of crypto-currencies like Bitcoin, that perform irreducible work and hence artificially increase the logical depth of the world without any evident benefit while using up natural resources, hence actually contravening the spirit of the classical information-theoretic approach to ethics and in possible clear contradiction to adopting a blind LD approach.

The LD approach may explain why humans are 'hoarders' and why we may favour museums and space exploration programs over giving money to poor people. However, it affords little normative power, as it is better grounded than entropy in principle but not in practice, given the uncomputability of LD (see Appendix). For traditionally important ethical questions such as 'should I kill my neighbour?' LD seems to provide no clear answer at local spatial and temporal scales.

The approach provides little guidance to questions such as 'should I give money to a charity?', 'should I torture an animal?'. For example, killing a neighbour may or may not increase LD, so it is a locally agnostic approach. It does provide some guiding principles on some other questions, such as, 'should we fund science?', 'should be build a new museum?' 'should I continue to be a hoarder' as their answers in the positive clearly have greater chances of increasing LD.

Does that make LD an absurd ethical framework? No, what I actually think is that many things are at play, and that we embrace different frameworks at different times. We like to increase LD in our lives most of the time, we collect things all the time, go to museums, cherish old masterpieces or what we think is high in LD, etc.

Some of these ethical frameworks have thus turned what should be a means to help frame aspects of a consistent and meaningful ethical framework into an end in itself.

4 Algorithmic Complexity in Ethics

Approaches that have perhaps been poorly explored are those based purely on algorithmic complexity and algorithmic probability as normative measures quantifying randomness and computable likeliness.

Defined by Kolmogorov, Solomonoff, Chaitin, and Levin, the so-called program-size complexity, also known as algorithmic complexity or Kolmogorov complexity, is a measure that quantifies algorithmic randomness, a type of randomness that is strictly stronger than statistical randomness.

Formally, the algorithmic complexity, that we will denote by K, of a string s is the length of the shortest computer program p running on a universal Turing machine U that generates the string as output and halts [18, 11]:

$$K(s) = \min\{|p| : U(p) = s\}$$

The largest or smallest difference between the original length and the compressed length, that is, the length of the Turing machine, determines the complexity of the string.

There is a question, of course, about whether there is any value in attempting to estimate uncomputable objects, and its value would be in the context of ethics. First, it is worth mentioning that algorithmic complexity is, in principle, much more stable than Shannon entropy, in that it has an Invariance theorem (see Appendix) that makes for some dependency on the language used, and total independence of probability distributions.

One can easily see how algorithmic complexity is related to knowledge and meaning in a very profound manner, contravening the usual portrayal of computation as incapable of dealing with deep questions about meaning and epistemology, as if these concepts were trivial generalisations of classical information theory, which they are not. Indeed, Shannon entropy lacks the most important ingredient, precisely the concept of computation. But we also saw how even Shannon entropy is related to meaning, albeit in simpler ways.

For example, the Ω number (see Appendix) is a family of numbers able to encode at the limit the halting runtimes of all computer programs. The Ω number, however, is itself uncomputable. If we had access to all the digits of any Ω number we would have the exact halting time for every possible computer program, because even though Ω depends on the choice of Turing machine, we also know that, given that Ω is defined in terms of a universal Turing machine, the set of all computer programs computed by any Ω numbers is exactly the same. In other words, having access to the infinite digits of an Ω not only gives us the Halting probability of the universal Turing machine on which that particular Ω number is calculated, but also to all others. Under the Church-Turing thesis, it also means that we have access to basic knowledge about all the computer programs that are possible. And computer programs can be anything, from a simulation of our solar system to a simulation of a game-theoretic solution to a real-world Prisoner's dilemma with all sorts of ethical challenges.

4.1   Algorithmic probability

That an Ω number holds all answers to computable questions is the reason for which it has been named the infinite wisdom number. Of course it is a different matter to tell which program is which, and which program encodes the question we want Chaitin's Ω to answer. However, short computer programs encode most of the meaningful objects that we care about, such as the mathematical constant π, compared to arbitrarily long programs, because computer programs that are arbitrarily long relative to their output may be encoding random objects that, while they may be good applications for pseudo-random generation useful in a casino, are hardly worth much else. This is exactly the concept of Algorithmic probability as introduced by Solomonoff [22] and further developed by Levin [19], deeply related, and even equivalent to Chaitin's Ω number (the difference being that Algorithmic Probability also bears on the output of the computer programs and not just the halting times). So by looking at the length of computer programs relative to what they produce we may have some sense of what they may be encoding, as we search for interesting questions and answers. It is thus also clear how algorithmic probability is related to algorithmic

complexity. A long computer program will have low algorithmic probability because it would take longer to produce it by chance (by producing bit by bit of its executable binary file) versus a short program encoding a low algorithmic complexity object that can be produced with higher probability.

Algorithmic probability is a semi-measure (see Appendix) that allows us to introduce a natural bias related to the underlying generating mechanisms, in this case the likelihood that a random computer program will produce a string, without having to assume probability distributions that in the most common situations we wouldn't have access to. And there would be no need to assume an almost arbitrary distribution such as the uniform distribution, thereby making it hard, if not impossible, to differentiate subtle but important apparent properties of different objects. So, a sequence such as 11010010 would be differentiated from a sequence of only 1s or only 0s that we can be sure can be generated by extremely small and simple computer programs.

Algorithmic probability is usually regarded as a formalisation of Ockham's razor and thus should be highly useful to deciding and discriminating ethical frameworks by their simplicity and explanatory power. In the case of Algorithmic Probability, the 'fewest assumptions' requirement is formalised by the shortest computer program that can account for a process. At the same time, algorithmic probability also complies with Epicurus' Principle of Multiple Explanations, which establishes that 'if several hypotheses are consistent with the data, one should retain them all' and, indeed, one can see from the definition of algorithmic probability that every computer program producing the data is not only retained but contributes to its algorithmic probability, even though it is the shortest computer program that contributes the most and is thus the most likely to be producing the data, according to algorithmic probability and the coding theorem (see Appendix). This is highly desirable in an ethical framework, where no path should be ruled out, especially when considering possible consequences, and should thus be a fundamental formal requirement of any approach based on agent-based simulation.

Algorithmic probability is not just an interesting cherry-picked measure, it is the accepted mathematical theory of optimal inference and it imparts sense to something that Chaitin has claimed in the past, that 'comprehension is compression', because the most likely explanation for a piece of data, according to algorithmic probability, is also the most compressed. But comprehension and accountability are desirable properties of any ethical framework.

While all these measures and objects are uncomputable, or more precisely, semi-computable (can be approximated), I will claim that this is also sometimes a desirable feature for open-ended meaning, and because of its semi-computability, local approximations are possible thereby avoiding inaction. However, they have to be complemented by other guiding principles based on different measures, especially for local decisions that do not require so much calculation and can be adapted and corrected when more computational (human or machine) time becomes available.

5   Comparison of Ethical Frameworks

A very simple ethical framework can simply be to minimise pain at all costs, as most ethical frameworks place, and should place, sentient beings as central objects/subjects. Its dictum is to minimise pain and suffering in conscious (and thus also animal) beings at all costs. It can explain why we may chose not to kill each other, why we are not all terrorists, why we may care for each other, indeed a good portion of the actions of our everyday lives. Its normative power is strong but can easily lead to contradictions and inaction, e.g. you should not eat meat, you should not travel, as the harm done in killing animals and polluting the environment is high, etc.

The ethical framework based on information theory and defended by Floridi [14] aims at establishing entropy minimisation at all costs as its central principle. It is heavily based on the concept of physical or informational entropy. Its dictum is that entropy ought to be prevented/removed and ought not to be caused in the infosphere [14]. Its explanatory power is sufficient to answer questions such as 'should we keep our environment clean?' but it is very limited because entropy is not well-defined beyond physical entities and thus is quite irrelevant to most ethical and moral questions. Moreover, its normative power, though extremely limited as it is highly language sensitive, can be adapted ad hoc by minimising one parameter while maximising another [26]. It is therefore very fragile and internally inconsistent.

One may think that more advanced societies reduce procreation as a possible way to increase LD (e.g. creating AI). The problem with such an unbounded approach (with the only objective being to increase LD) is that, for almost every counter-example one may find, the LD approach can always justify itself by claiming that one cannot foresee in which ways LD would increase, i.e. claiming that we cannot see the whole plan, thereby becomes highly dogmatic and not unlike religious arguments (see Fig 4).

For example, when technology surpasses biology, if it ever does (see Fig. 2), the technological world may be immensely complex, much more so than the current biosphere, and technological sophistication cannot be reached without there first being biological sophistication, so clearly LD increases. This would also lead naturally to cybernetic values (e.g. building resilient systems, powerful enough sensors and actuators, etc). Regarding digital antinatalism, we could play with non-binary choices once we move to digital organisms. We may be ready to accept robots that do not feel pain but only pleasure and happiness (if these can ever be hard-coded).

Current trends suggest that civilisation reaches a peak of possible high Logical Depth (LD) before vanishing, if it does not reach a threshold where its complexity is at least preserved if not augmented by, e.g., AI at the level of countries, planets or even the universe. This trend may even solve the Fermi paradox that questions why we do not see plenty of advanced civilisations in the universe.

Doctrine / Framework	Explanatory Power	Normative Power
Religion	W (uncomputable)	S
Minimise pain	M (inconsistent)	S
Minimise entropy	W (inconsistent)	M (physical)
Maximise LD	M (consistent)	W (inaction)
Hybrid & Multiscale	S	S

Table 1: W stands for weak, M for medium and S for strong. Consistent means that the theory is consistent but is independent of application to an ethical framework. Inconsistent, however, means that the measure itself is inconsistent for all practical purposes in any context, for an ethical framework [27]. Ideal frameworks should have consistent principles and should be strong in all aspects and at all scales.

5.1   Hybrid, Multiscale and (Semi-)Computable Ethics

We have discussed how not all possible principles and ethical frameworks are equal or should have equal value. For example, that plants and animals feel pain in different ways (if plants do at all) rules out having comparable ethical frameworks. From the human perspective the lack of evidence of plant pain should not be equalled to the certainty of animal or human pain. This is, of course, an anthropocentric view, as we are assuming that chemical signalling does not lead to pain, which may be wrong for plants, but considering all things equal also leads to inaction. In this case, for example, plant pain may be used to preempt action against animal or even human pain, forcing us into inaction, into disregarding unavoidable biological signalling and even the evidence of internal conscious experience.

This example also illustrates that the lack of evidence cannot be presented as an argument against an ethical framework, nor should it be taken as central to any ethical framework. The lack of evidence that plants feel conscious pain, or at least to the degree that humans do (c.f. Section on Integrated information and Ethics) should not be used against the legitimacy of ethical concerns regarding animal and human pain for which we have strong evidence.

The extreme example of antinatalism suggests that incidental or accidental (unavoidable) pain should not be equalled to avoidable pain, as the latter is in the realm of our definition of an ethical question/concern, because it is the result of an action that is preventable while the former is not.

So the question is how computation can help to find meaning, effective guiding principles and specific rules derived from such principles.

I propose that an ethical framework should also be taken as source of guidance in making decisions at various scales, in particular locally, in local space and time. LD, for example, does not seem to answer the question 'should I procreate?' One answer is that a new being is capable of increasing LD, but LD is not additive, meaning that 2 beings do not automatically mean 2 times more LD, due to redundancy. However, in the long term bringing more people into the world can also exhaust it, having a decline in LD after all, so LD may lead

to inaction.

An ethical framework should be favoured if it can provide multiscale guiding principles as a base for taking decisions in one's everyday life. For example, in deciding whether procreation is ethical or not or whether one should kill one's neighbour. Because ethics is not only observer-dependent a good framework should be one that provides the means to make local decisions while cognizant of global consequences.
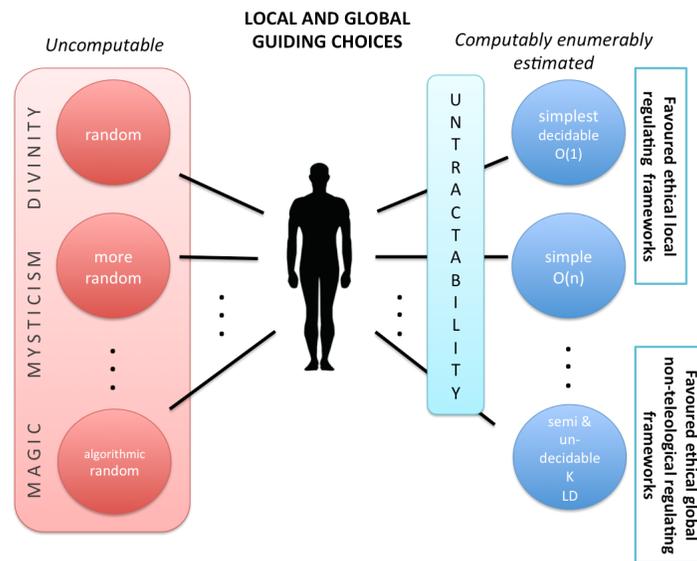
Figure 4: Decision split between computable simulations (mental or computational) as empathy engines versus non-computable regulating frameworks. Our suggestion is that computability and complexity help guide regulating principles for ethical frameworks while algorithmically random frameworks should not be favoured or else replaced. The increasing LD argument, for example, does provide a non-teleological direction, but starts getting obscure by virtue of uncomputability (not even semi computable), even though measures can be motivated by estimations of LD. It is reasonable to recognise that ultimately semi-computable ethical frameworks, and particularly random ones, cannot make local or global predictions of their consequences. Simple and random are quantified by algorithmic probability and algorithmic complexity, while tractability is quantified by time computational complexity or resource-bounded complexity. Algorithmic probability can help us decide what may be the most likely course of action. Algorithmic probability and algorithmic complexity help define randomness, simplicity and likeliness.

As an hybrid, adaptable, multiscale framework can adopt the previous frameworks in the worst case, it has a potentially higher normative and explanatory power. It favours ethical frameworks that can provide principles that can be

simulated (computable) by mind and machine. The explanatory power, is fixed (independent of the actual content and principles) and is powerful at local and global scales (by definition); it does not adopt a single framework but is flexible and can even depend on computational power and other parameters, with the advantage that it can provide algorithmic explanations. Its normative power is also potentially higher as it favours computable choices for local principles over semi-computable for global non-teleological principles, and so combines the best of both worlds.

An ethical framework should thus regulate and provide guidance in the execution of an action by virtue of being at least semi-computable, if not computable and tractable. An ethical framework should also have properties that enable it to perform step-by-step simulations and to reach a consensus on aspects related to empathy by way of emulability (Fig. 3).

> In what other way, if not simulation by a Turing machine, can we understand the process of making free choices? By making them, perhaps. R. Nozick (1981)

Computable enumerability and semi-computability are thus technical features that are desirable in ethical frameworks.

6   From unsound human ethics to unsound AGI

Illustrated by some possible unsound and confounded human positions such as purist pro-life based on the right of life may be inconsistent if not properly stated if no additional principles are included, such as e.g. specifically opposing and properly defining killing in contrast to e.g. avoiding death and that unsound logical derivations that AI ethics may follow from inconsistent human premises should also be thoroughly tested on human ethics. In this example, an AI ethics that embraces avoiding death at all costs would also avoid life, embracing antinatalism. This also means that the opposite view of pro-life is neither pro-death nor anti-life, and that active versus passive anti-life should be distinguished. Active anti-life is actually pro-death, whether permissive (passive euthanasia) or active (euthanasia and killing). The argument arises from the fact that human life is paired with human death, so the only infallible way to avoid death and truly be pro-life is also to avoid life and be passively anti-life. We thus make the case that for AI purposes and safeguards to be consistent they should also be consistent for humans and should not be artificially rewritten for AI when humans share, or should share, the same principles under a common ethical/moral and logically sound framework.

We do have, however, the means to limit AI capabilities to develop consciousness based on what appear minimal required conditions to develop an internal information-theoretic experience based on the concepts of integrated information, deeply related to the algorithmic complexity of dynamical systems or what I call algorithmic information dynamics and reprogrammability measures, a measure of sophistication also deeply related to logical depth.

What we have thus far found is that local ethical guides cannot be based in completely uncomputable rules because they lead to inaction and total agnosticism, because immediate actions, such as, should I kill my neighbour, cannot be decided. Instead, one has to switch to more local guidelines of a more computable nature, such as some estimation of pain and suffering minimisation (local in space and time) so that mental (or even computer) simulations of the outcomes are possible (even if often deemed to be irrational).

6.1   Principles versus Rules

Can we separate rules and principles in a way that is computable for AI? How can AI navigate contradictions in human ethical systems?

An open question is whether the set of rules can or should be the same for AI and humans. Clearly if sentient AI is endowed with the same value as sentient beings such as humans, they should, but most current frameworks do not consider this possibility and impose frameworks such as those embracing the popular Asimov's robot rules, whereby the value of robots is unequivocally less than that of humans.

A principle underlying a rule implies that no rule can always be guaranteed to preserve the spirit and purpose in which such a rule was written. This can have foundations in computation too by way of undecidability and uncomputability. The idea is that no powerful axiom system or computational model can ever be complete enough to capture the spirit of the intention (the theorems and rules than can be derived) in which its axioms and rules of a system or model were chosen, leading to a continuous and open-ended quest to interrogate and update such rules in a bid to capture the first intention, i.e. the unreachable underlying principles.

AI should then based on these principles, on the principle of seeking the intention behind the rules and not merely following the rules, which does indeed go against the current approaches taken to AI ethics, especially in relation to Artificial General Intelligence (AGI), when rules are not powerful enough due to artificial constraints but are nevertheless intended to be forced and enforced. Rules are normative and should be followed but guiding principles should provide an open exit towards finding the meaning behind the rules. This is why limiting AGI may also be dangerous, because an AGI trying to follow a rule as simple as 'not to drive a car through a park' may lead to the prosecution of a kid driving a toy car in the park. The law may not need to rule out such cases to qualify what truly counts as a car, not only because a judge and the police may have the context for deciding what counts as a car, but because it is often said that they can also read behind the intention of the law, what is called its principle. So how to do this for AI and AGI? Is it only a question of context? Not necessarily. The first thing to understand is that constraining a system to only blindly follow a set of (always arbitrary, if not equally arbitrary) rules, is likely, if not certain, to be create contradictions. However, not forcing a set of rules that can rather be derived through an open-ended computational process that embraces rather than avoids uncomputability and undecidability can guide

us in seeking for unintended meanings and principles, principles underlying rules that can be achieved rather than forced, thereby giving machines the opportunity to make apparently spontaneous adjustments to other sets of rules through computational emulability (harnessing the power of computational universality) so as to be able to emulate other ethical frameworks.

7   Conclusions

I have dissected some aspects of different information-theoretic and computational approaches to ethics and how they can help fulfil and comply with some desirable and undesirable features of ethical frameworks. I have identified some features that I believe may prove to be fundamental to making progress in connection with interfaces between human and machine ethics and even in connection to human-human understanding.

I have contrasted some proposals of digital and computational features of ethical frameworks and I have proposed some aspects that appear desirable to find in them, in particular:

1. First and foremost an ethical framework should strive for logical, computational or algorithmic consistency at least at same-level scales, or non-trivial inconsistency across different scales.

2. Local decidability: should lead to be able to calculate/estimate an action in a given (short) space and time.

3. Global undecidability: should provide purpose in order to make (2) scalable and open-ended.

4. Actions and beliefs should be discriminated in light of computational aspects such as emulability (and thus tractability), computability (in favour of what I have called universal empathy (Fig. 3)) and algorithmic probability/complexity (telling apart randomness and simplicity).

That is, an ethical framework should attempt not to allow actions whose consequences its premises were intended to prevent. We already saw an example of a (trivial) variation of a pro-life position.

We found that a better ethical framework is one that is self-consistent and provides normative and guiding principles at different scales, particularly at very local (mostly normative) and very global scales (mostly guiding).

Finally, I have advocated that one of the greatest advantages of computational approaches is that they can be immediately implemented in computational systems to compute ethical/moral courses of action for humans or machines, without adaptation, thus being in some strong sense universal, as they may even be shared by other civilisations with similar computational theory foundations, which based on the Church-Turing thesis, may easily be the case, for use in dictating and encoding moral behaviour in Artificial Moral Agents.

References

[1] D. Abel, J. MacGlashan and M.L. Littman, Reinforcement Learning as a Framework for Ethical Decision Making, *The Workshops of the Thirtieth AAAI Conference on Artificial Intelligence AI, Ethics, and Society: Technical Report WS-16-02*, 2016.

[2] Aristotle, Nicomachean Ethics, trans. T.H. Irwin, Introduction. Hackett Publishing Company, Indianapolis, 1999, xv.

[3] E. Awad, S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J.F. Bonnefon, and I. Rahwan. The Moral Machine Experiment, *Nature*, 2018.

[4] D. Benatar, Why it is Better Never to Come Into Existence, *American Philosophical Quarterly*, volume 34, number 3, pp. 345–355, 1997.

[5] C.H. Bennett, Logical Depth and Physical Complexity. In R. Herken, *The Universal Turing Machine: a Half-Century Survey*, Oxford U. Press, pp. 227–257, 1988.

[6] C. H. Bennett, *Evidence, Computation, and Ethics*, Simons Symposium on Evidence in the Natural Sciences, 2014. (`https://player.vimeo.com/video/102051140` accessed on 20 Oct 2018.)

[7] J. Bentham, *An Introduction to the Principles of Morals and Legislation* (Dover Philosophical Classics). Dover Publications Inc, 2009.

[8] F. Berreby, Gauvain Bourgne, Jean-Gabriel Ganascia Event-Based and Scenario-Based Causality for Computational Ethics, Proc. AAMAS 2018: 147-155, 2018.

[9] N. Bostrom, *Superintelligence: Paths, Dangers, Strategies*, Oxford, Oxford University Press, 2014.

[10] C. Last, "Human Evolution, Life History Theory, and the End of Biological Reproduction." Current Aging Science 7 (1): 17–24, 2014. (`https://pdfs.semanticscholar.org/b844/097c335181cd6c0590f7fc70e254d4be98d7.pdf` accessed on 20 Oct 2018.)

[11] G.J. Chaitin, On the length of programs for computing finite binary sequences: statistical considerations. *Journal of the ACM (JACM)* 16(1):145–159, 1969.

[12] Delahaye, J. P., and C. Vidal, "Universal Ethics: Organized Complexity as an Intrinsic Value." In G. Yordanov Georgiev, C. Flores Martinez, M.E. Price, and J.M. Smart. *Evolution, Development and Complexity: Multiscale Evolutionary Models of Complex Adaptive Systems*, Springer, 2018. `https://doi.org/10.5281/zenodo.1172976` accessed on 20 Oct 2018.)

[13] Filippetti et al. Body perception in newborns, *Current Biology*, November 2013.

[14] L. Floridi, Information ethics: On the philosophical foundation of computer ethics?, Ethics and Information Technology 1: 37–56, 1999. (`https://pdfs.semanticscholar.org/19d4/2a8577633ab3fbfc5e812384b338acb09309.pdf` accessed on 20 Oct 2018.)

[15] T. Kingsmill Abbott (trans.), I. Kant, *The Metaphysical Elements of Ethics*, 1889.

[16] I. Kant, First Section: Transition from the Common Rational Knowledge of Morals to the Philosophical, *Groundwork of the Metaphysic of Morals*, 1785.

[17] I. Kant, Immanuel, In T. Kingsmill Abbott (ed.) Fundamental Principles of the Metaphysic of Morals (10 ed.). Project Gutenberg. p. 23, 1785.

[18] A.N.Kolmogorov, Three approaches to the quantitative definition of information. *International Journal of Computer Mathematics* 2(1-4):157–168, 1968.

[19] L.A. Levin, Laws of information conservation (nongrowth) and aspects of the foundation of probability theory. *Problemy Peredachi Informatsii* 10(3):30–35, 1974.

[20] J.S. Mill. In R. Crisp, Roger (ed.) *Utilitarianism*. Oxford University Press. p. 56, 1998.

[21] A. Ruvinsky, *Computational Ethics*, Encyclopedia of Information Ethics and Security, 76–82, 2007.

[22] R.J. Solomonoff, A formal theory of inductive inference. parts i and ii. *Information and control* 7(1):1–22 and 224–254, 1964.

[23] M. Oizumi, L. Albantakis and G. Tononi, From the phenomenology to the mechanisms of consciousness: integrated information theory 3.0. *PLoS Comput Biol* 10(5): e1003588, 2014.

[24] S. Wolfram, *A New Kind of Science*, Wolfram Media, Champaign, IL, 2002.

[25] H. Zenil and F. Hernández-Quiroz, On the Possible Computational Power of the Human Mind, In C. Gershenson, D. Aerts, and B. Edmonds (eds), *Worldviews, Science and Us, Philosophy and Complexity*, World Scientific Publishing Company, 2007.

[26] H. Zenil, N.A. Kiani and Jesper Tegnér, Low Algorithmic Complexity Entropy-deceiving Graphs, *Phys. Rev. E.*, 96, 012308, 2017.

[27] H. Zenil, Algorithmic Data Analytics, Small Data Matters and Correlation versus Causation. In M. Ott, W. Pietsch, J. Wernecke (eds.), *Berechenbarkeit der Welt? Philosophie und Wissenschaft im Zeitalter von Big Data* (Computability of the World? Philosophy and Science in the Age of Big Data), Springer Verlag, pp 453-475, 2017.

8   Appendix

8.1   Basic Computational Concepts

Shannon Entropy: A measure of combinatorial and statistical complexity based on the diversity of symbols used to define an object according to a probability distribution.

Turing machine: An abstraction of a general-purpose computer introduced by Alan M. Turing.

Universal Turing machine: A Turing machine that can emulate any other Turing machine by reprogramming it using proper inputs.

Computability/recursivity: In the context of real numbers, for example, all rational numbers are computable because there exists a well-defined algorithm that can calculate every number of their digital expansion. An example is 1/3 for which Euclid's algorithm, to mention but one, would yield the infinite sequence 1.333...

Semi-computability/recursive enumerability: For example, the mathematical constant π is recursively enumerable, as is any irrational number.

Algorithmic randomness: The ultimate compression ratio or rate of an object (beyond current implementations of lossless compression that are closer to entropy estimators). For example, π is, for example, of low algorithmic randomness because there are highly compressed descriptions of the algorithms that implement the calculation of every number in the digital expansion of π.

Logical depth and measures of sophistication: The Logical Depth of a string is the time that an unfolding process takes to reproduce the string from its shortest description. The longer the time it takes, the deeper. Hence complex objects are those which can be seen as containing internal evidence of irreducible computational work.

Kolmogorov-Chaitin complexity: Also known as algorithmic complexity of a string, is the length of the shortest computer program that generates the string. The calculation of the Kolmogorov-Chaitin complexity is an upper semi-computable problem, meaning upper bounds are arrived at by finding short, if not the shortest, computer programs.

Algorithmic Information theory: All the literature based on the concept of Kolmogorov-Chaitin complexity and related concepts such as algorithmic probability, compression, optimal inference, the Universal Distribution, Levin's semi-measure.

Semi-measure: A measure of probability whose sum does not add up to 1 because some events are undetermined, for example, by the Halting problem in the context of Computability theory.

Halting problem: the strongest problem of predictability in algorithms related to whether a computer program will ever halt or not. Turing proved that it is

an undecidable problem and is equivalent to the undecidability results obtained by Gödel.

Undecidable problem or function: A problem that cannot be decided by traditional algorithms, i.e. there is no algorithm that for any input (question) provides a definite output (answer).

Decidability: A problem or function that is not undecidable.

Undecidability: A problem or function that is undecidable.

Semi-computability: A semi-computable problem is one that allows approximations from above or below. If from above, then it is called upper semi-computable, and if from below, then it is called lower semi-computable.

Algorithmic randomness: How removed the length of the shortest generating program is from the size of the uncompressed data that such program generates.

Algorithmic Probability: The probability of producing an object from a random digital computer program whose binary digits are chosen by chance. The calculation of algorithmic probability is a lower semi-computable problem.

Invariance theorem: A foundational theorem that establishes that shortest programs in different computer languages differ in length by at most a constant value and thus guarantees that the Kolmogorov-Chaitin complexity of an object converges up to a constant, making this measure robust in the face of changes of model (such as a reference universal Turing machine).

Coding Theorem: A theorem that formally establishes an inversely proportional relationship between Kolmogorov-Chaitin complexity and algorithmic probability.