Summer 8-5-2023

Proposing a Measure of Ethicality for Humans and AI

Alejandro Jorge Napolitano Jawerbaum
Duquesne University

PROPOSING A MEASURE OF ETHICALITY FOR HUMANS AND AI

A Thesis

Submitted to the McAnulty College and Graduate School of Liberal Arts

Duquesne University

In partial fulfillment of the requirements for

the degree of Master of Science

By

Alejandro Jorge Napolitano Jawerbaum

August 2023

PROPOSING A MEASURE OF ETHICALITY FOR HUMANS AND AI

By

Alejandro Jorge Napolitano Jawerbaum

Approved June 19, 2023

Dr. Adam Drozdek
Associate Professor of Computer Science
(Committee Chair)

Dr. Lauren Sugden
Assistant Professor of Statistics
(Committee Member)

Dr. Patrick Juola
Professor of Computer Science, Cybersecurity Studies Coordinator, Joseph A. Lauritis, C.S.Sp. Endowed Chair in Teaching and Technology
(Committee Member)

James Swindal
Professor, Henry Koren, C.S.Sp Endowed Chair in Scholarly Excellence, and Director of Undergraduate Studies
(Committee Member)

Dr. Kristine Blair
Dean, McAnulty College of Liberal Arts

Dr. Karl Wimmer
Department Chair and Associate Professor, Mathematics and Computer Science

ABSTRACT

PROPOSING A MEASURE OF ETHICALITY FOR HUMANS AND AI

By

Alejandro Jorge Napolitano Jawerbaum

August 2023

Thesis supervised by Dr. Adam Drozdek

Smarter people or intelligent machines are able to make more accurate inferences about their environment and other agents more efficiently than less intelligent agents. Formally: 'Intelligence measures an agent's ability to achieve goals in a wide range of environments.' (Legg, 2008)

In this dissertation we extend this definition to include ethical behaviour and we will offer a mathematical formalism and a way to estimate how ethical an action is or will be, both for a human and for a computer, by calculating the expected values of random variables. Formally, we propose the following measure of ethicality, which is computable, or at least can be approximated: The ethicality of an action can be measured as the change in ability to reach a goal of all agents affected by an action, for each action taken, and weighted by the intelligence of the actor.

Thus, we claim that agents cannot be morally responsible for consequences they are not smart enough to infer. They are (morally) responsible only for what they could foresee given their intelligence. Finally, intelligence provides us with estimated and exact measures: The former is used by the actor who must pick a course of

action, especially in the heat of the moment which would require a lot of estimations to be made, whereas the latter can be used by any agent with sufficient computational power and time and should deterministically yield the same results.

ACKNOWLEDGEMENT

With special thanks to Drs. Adam Drozdek and Patrick Juola, for their patience and thoroughness.

TABLE OF CONTENTS

Chapter 1. Introduction

We can use a person's ability to get what they want, irrespective of what else is going on, as a measure of how smart they are. This measure can be weighted by the complexity of the environment (or environments) they are acting in, to account for differences in the challenges different people face while pursuing a goal, in order to get a normalized, universal measure of intelligence. Herein we propose that the ethicality of our actions can be measured using intelligence as a basis: As we try to get what we want, we make decisions on how we will impact others along the way, and smarter people are able to predict the extent of that impact better. In other words, smart people are able to make more accurate inferences about their environment and other agents more efficiently than people who are not as smart. We must therefore account for how intelligent someone is when assessing how ethical or unethical their actions were: By understanding how intelligent actors are, we can estimate how well they will be able to predict and infer the impact of their actions on others, and compare that to the actual objective outcome in order to understand their moral responsibility.

Thus, we claim that actors cannot be morally responsible for consequences they are not smart enough to infer —though they can still be legally, financially, or socially responsible of course; they are (morally) responsible only for what they could foresee.

But in order to get what we want, we have to *act*. Thus, we call agents, human or otherwise, performing actions *actors* and, by definition, actors have *agency*, which is not the same as freedom. We say something has agency when it is able to operate with autonomy, meaning that it has the capability of acting by itself. When talking about ethics, however, we must be more specific: Moral agency is having the capacity to make informed and un-coerced decisions in reference to some notion of right or wrong. With humans, this right or wrong comes as a complex interaction of social, psychological, and biological factors.

One of the main issues we encounter is the problem of enumerating values. Take

Aristotelian ethics as an example: By enumerating a set of virtues, he is being directly normative, telling us how to behave. AI, however, is like a genie in the bottle; it carries out our orders to the letter, and if something is not specified the consequences could be potentially harmful, especially if this AI is making decisions that impact people. This is called the **value-loading problem**: How do we design an intelligence that wants high-value, long-term, outcomes beneficial for all intelligent life? If we assign or hard-code a list of values, an off-by-one error, meaning to miss or incorrectly formalize a necessary value, or to add an incorrect value, can be potentially catastrophic in the case of a superintelligent machine, because it could lead to missing a key value that would have prevented, for example, drugging us all with heroin in order to maximize happiness.

A step in solving this problem is called **indirect normativity**. In Bostrom's (2014, p. 163) words, indirect normativity is a "process for deriving a standard" with which to judge possible values and interactions. Kantianism is a simple example of this, the indirectly normative statement being that one should act in a way such that the principle of non-contradiction —however, for Kant, this is based on the principle of treating every person as an end in itself. A **goal**, or an objective, can be predetermined by a programmer, biology, or a whim of the actor, and these goals are independent of intelligence as per the **Orthogonality Thesis**, which states that "intelligence and final goals are independent variables: any level of intelligence could be combined with any final goal"(Bostrom, op.cit. p.105). This can be expanded into ethics: Being ethical, after all, is a goal regardless of how it comes up in the decision process. A higher intelligence merely predicts *how efficiently* we work towards our stated goal in the given environment, but does not have an impact on whether we choose to be ethical, morally neutral, or unethical. On the other hand, even if an agent does not possess the same *final* goals as others, the **Instrumental Convergence Thesis** "holds that superintelligent agents having any of a wide range of final

goals will nevertheless pursue similar intermediary goals because they have common instrumental reasons to do so." (Bostrom, ibid.).

Legg (2008) defined the possible rewards for an agent as "A signal that indicates how good the agent's current situation is" (p.72). What 'good' means depends on the specifics of each actor and its causal history. A reward is basically whatever an algorithm might "want" (e.g. its reward pathways, seen as optimization of what it was programmed to do, or for humans an evolutionary selection process), and what grants us a sense of accomplishment, or the feeling that comes with happy chemicals in our brain. Note, moreover, that a reward is a *signal*, not an object or a concrete thing, but rather a proxy for the accomplishment of or progress towards a goal. Take hunger as an example: Eating a hamburger is not the reward for satisfying hunger, it is the means to do so, but the actual reward is the endorphins released when you eat, or more precisely the feeling that they cause. As a matter of fact, the release of endorphins that comes with eating is independent of the pleasure we may feel by eating (Tuulari, Tuominen, et. al., 2017), meaning the pleasure that comes from satisfying hunger and the one that comes from the taste of the food, are two different reward-signals.

When it comes to pleasure from the particular food, we are hard-wired to prefer fat, sugar, and salt: Evolutionarily, these things used to be rare enough that we now have reward mechanisms around consuming them. Fast food exploits these mechanisms by providing some combination of the three with relatively low nutritional value. This provides immediate gratification in the form of chemical rewards, which brings medium to long term problems, as opposed to diet and exercise which are difficult at the beginning and bring longer lasting and stronger effects in the long term. This preference of short-term rewards at the cost of trivial or nonexistent improvement (rather than medium- to long-term, more significant rewards) is called **wireheading.** This behaviour takes place in reinforcement paradigms, including be-

havioural reinforcement in humans as in the fat, sugar, and salt example. Another example are AIs that receive some sort of reward through improvement. The general idea, following the example of self-improvement, is that the AI will follow a path towards reward-maximization, rather than goal-achievement, in a way analogous to the behaviour of addicts. It may, for example, engage in trivial improvements which enhance one of its aspects at the cost of another, such that it may continually be rewarded for making such a change and for later undoing it and gaining back what it had lost.

The second reward, independent from the first, is the feelings caused by endorphins from the fact of having eaten, regardless of what we ate. This is the satiety reward, and it comes with having achieved a goal (that of not starving).

Furthermore, keep in mind that rewards can be positive or negative. A negative reward is merely a signal that we are not doing well, such as in reinforcement learning, where the algorithm has to reach a target "points", but each time unit elapsed could, for example, result in a point being subtracted. In the human case, one of these signals is pain. We can also consider shame, guilt, and other more abstract signals as negative rewards.

How ethical (or unethical) an actor can decide to be, should they choose to act at all, depends on how smart they are: If they choose to be ethical, smarter actors can more consistently maximize any measure of ethicality due to their inferential capabilities, whereas someone with lower intelligence will have more unexpected outcomes. Given an actor's intelligence, we can estimate how well they will be able to predict and infer the impact of their actions on other agents. The dark side of this is when agents choose to be unethical; the smarter they are, the worse they can act. Thus, the ethicality of our actions can be measured using our intelligence and the reward functions of those our actions affect: These can be estimated by inference —deduction and induction for humans— and the best way to get information for that inference

is to discuss our possible actions with other agents. It bares repeating that one does not have to be "smart" to be ethical. Instead, smarter agents can, if they have the goal to be ethical, maximize the ethicality function more efficiently.

Thus, intelligence provides us with both an estimated or heuristic, and an exact measure of ethicality: The former is used by the actor, human or otherwise, who must pick a course of action, especially in the heat of the moment which would require a lot of estimations to be made, whereas the latter can be used by any agent with sufficient computational power and time, and should deterministically yield the same results no matter who calculates it.

1.1 A formalized definition of intelligence

Legg (2008) formalized a universal measure of intelligence, starting from the following definition: 'Intelligence measures an agent's ability to achieve goals in a wide range of environments.'

Legg then defines the *expected future discounted rewards* V for some agent π interacting with an environment μ (thus we call them V_μ^π), with a set of discounts applied for temporal preference. The reason for discounts is not only as weights, but because of the difficulties in predicting the future. This is the issue of the time value of money, for example: Having money in the present is, on average, worth more than having the same amount in the future due to its earning potential; delayed investment is a lost opportunity, not to mention inflation. In the case of reward signals r, if one can secure a positive reward now, then it is preferable than having it later because of the uncertainty of the future. Likewise, the avoidance and postponement of negative rewards is preferable, rather than suffering the consequences immediately. In this case, the weighting normalizes rewards such that the sum over all rewards is finite, and it "weights the reward at different points in the future which in effect defines a temporal preference" (Legg, ibid).

Discounting normalizes rewards to make their sum finite, and weights them ac-

cording to temporal preferences. A reward summable environment $\mathbb{E}$ can implicitly factor in time discounts by bounding the rewards, $r \in [-1, 1] \cap \mathbb{Q}$ —thus rewards are discrete and come from a countably infinite interval— such that $\underset{\mu \in \mathbb{E}}{\forall}$ "the expected value of the sum of rewards is also finite and thus discounting is no longer required." Note that temporal preferences are not ignored, but are rather included in the environment. For the purposes of this work, we can consider rewards to be discrete random variables: While current AI is "single-minded," in the sense that it seeks to maximize or minimize a function that's hard-coded into it, more complex AI may have to juggle several reward functions depending on its goals. For a system as complex as humans, each reward can come from a different black box or distribution; rewards then become an empirical observation from each distribution, and can be treated as a probabilistic proxy for goal achievement.

E is the notation that represents the *expected value*, which, for a discrete random variable with countably infinite outcomes such as rewards, can be conceptualized as the weighted sum of reward values, weighted by the probability of each value.

Thus we have

$$V_\mu^\pi := \mathbf{E}\left(\sum_{i=1}^{\infty} r_i\right) \leq 1 \text{(p.73)}$$

The above equation can be rewritten in the following manner for some $x \in [-1, 1] \cap \mathbb{Q}$:

$$V_\mu^\pi = \mathbf{E}\left(\sum_{i=1}^{\infty} r_i\right) = \sum_{i=1}^{\infty} \mathbf{E}\left(r_i\right) = \sum_{i=1}^{\infty} \sum_{x \in [-1,1] \cap \mathbb{Q}} x P(r_i = x)$$

Given the Kolmogorov complexity of an environment $K(\mu)$ such that an Occam's razor term can be defined as $2^{-K(\mu)}$ in order to decrease the weight of an agent's performance proportionally to the complexity of the environment, the formal definition of "the *universal intelligence* of an agent π is its expected performance with respect to the universal distribution $2^{-K(\mu)}$ over the space of all computable reward-summable

environments $\mathbb{E}$, that is," (p.77)

$$\Upsilon(\pi) := \sum_{\mu \in \mathbb{E}} 2^{-K(\mu)} V_\mu^\pi$$

It is worth noting that, since $V_\mu^\pi \leq 1$ —which implies that negative rewards can sum up to -1— and $2^{-K(\mu)}$ can only result in values in $(0, 1]$, $\Upsilon(\pi)$ is also bounded in [-1,1].

The equation above holds with the caveat that "The main drawback is that the Kolmogorov complexity function K is not computable and can only be approximated" (ibid). Nonetheless, anything that can be digitized can be compressed, and therefore we can approximate it fairly accurately.

Note that this holds (viz. it is valid, informative, dynamic, general, based on Turing computation, etc.) with some qualifications:

(1) Intelligence is dependent on the level of specialization, causal history, and reward-systems of an agent. Reward-systems is our term, involving factors such as prioritizing a forward-looking set of goals, meaning investing in larger future rewards rather than simply maximizing the next one (aka deferral of gratification).

(2) Legg claims it is impractical, because the Kolmogorov complexity function is not computable. This, too, will hold for our own proposal of ethics, at the time being.

(2.1) It is a definition rather than a test.

(2.2) However, approximations are possible, and we consider this to be good enough for now.

(3) "Reward is an interpretation of the state of the environment."

"In humans, this interpretation is internal" (89). We can also measure external reward as the outcome for some change in the utility function, which

would permit rewards to be negative.

(4) Searle's Chinese room experiment does not factor in, because understanding does not have measurable effects on outcomes (90). We will explore Searle's thought experiment in more depth later.

(5) Legg admits that, if consciousness, creativity, imagination, and so on have measurable effects, his Universal Intelligence measure is also a partial measure for these. However, we will later argue that, as latent constructs rather than explicit variables, they can only be intrinsically posited, but never externally verified, and as such should not be considered.

(6) In Legg's setup "the agent cannot decide what its primary goal is. It simply tries to maximize the reward signal defined by the environment." (92) This holds for machines in the present day, but not necessarily for humans or future machines: Given the orthogonality thesis, which shows intelligence is divorced from the choice of goals, we might consider reward-maximization to be goal-dependent rather than a goal in itself. Consider the human paradigm: One can choose a myriad of jobs, and the gratification obtained by them does not have a one to one relationship with the direct reward of working (paycheck). Also consider austerity; some of us enjoy such a thing, and need not maximize rewards to be satisfied, such that our intelligence could not be measured by those forms of rewards, but rather, for example, by the outcome of our intellectual pursuits, which yield no tangible reward other than internal gratification.

(7) Though the idea is not to create artificial humans, goal-selection factors a part in Artificial General Intelligence (AGI) theory insofar as it is assumed that the more intelligent an AGI, the more routes it will have at its disposal to get to a goal, even if hard-coded. Regardless, for this measure of intelligence goal-selection is important; there must be some kind of agency, and this will

become clearer soon, as we investigate the ethical implications of what has been mentioned, and the impact of intelligence on each ethical system.

(8) As for the Wolpert and Macready (1997) No Free Lunch theorem, we run into a complicated situation: Legg claims it does not apply to universal intelligence because "we have not taken a uniform distribution over the space of environments." This is problematic, because it means that we have no idea what kind of distribution we have, and so his claim is essentially that the theorem cannot hold because we do not know whether or not its hypothesis applies. Furthermore, there are theoretical agents such as AIXI (Hutter 2001a,b 2005, 2007) that are incomputable. Note that it is unknown whether humans have an incomputable aspect to our intelligence, but we are known to be highly flawed, inconsistent, and variable even within-subject, to the point where we cannot use ourselves as a baseline for any model of AGI.

Chapter 2. Our proposition to measure the ethicality of a decision

2.1 Formally

We propose the following measure of ethicality, which is computable, or at least can be approximated: *The ethicality of a course of action can be measured as the change in ability to reach a goal of all agents affected by an action, for each action taken, for each environment acted in, and weighted by the intelligence of the actor.* An action or proposed action is more ethical than another if it makes a higher nonnegative impact than the other. Our actions impact the environment or environments we participate in and intelligence of others, giving weight to the importance of Duties of Perfect Obligation, which for Kant are something all agents must abide by to be considered as behaving ethically. More precisely, duties of perfect obligations are what agents should *refrain from doing* since these duties are negative statements and violating them would result in a categorically unethical and non-universalizable course of action that will visibly impact other agents (Kant, 1797).

Given our considerations in the previous section, it could be well argued that other agents are included within the environments necessary for intelligence to be determined. However, this assumes a non-solipsistic AI, whereas modern AI are purely solipsistic, meaning it is not aware of other agents as agents. We will posit the ethicality of an action or course of action $E(A)$ as based on the expected value of a random variable. The set A may be composed of an individual action α or several $\alpha \in A$. To measure the ethicality when an actor chooses to be unethical, or can infer that its actions will produce an unethical result, the ethicality of its actions will be such that $E(A) < 0$. Given a set of agents Π, composed of the agents that the actor *can infer will be affected by its actions* and their intelligence functions $\Upsilon(\pi)$ *as inferred by the actor*, let π_0 denote the potential (or factual) actor who is considering or has already carried out a series of actions (A) and the individual actions $\alpha \in A$, and finally let the differential ΔV_μ^π denote the change in the expected future discounted

reward function of a particular agent an action under consideration would take in a given environment μ. Since $V_\mu^\pi \in \mathbb{Q} \cap [-1, 1]$, then the differential will be bounded between $[-2, 2]$.

Note that we will not be including the decision-maker itself within the sum. This helps avoid the issue of the actor benefiting itself by negatively impacting other agents such that the impact on others is less than the benefit the actor accrues.

2.2 Before the course of action is implemented

The actor itself, without the knowledge of its own intelligence, could infer the likely ethicality of its action by considering

$$\mathbf{E}(E(A|\text{knowledge}_{\pi_0}(\Pi))) = \sum_{\alpha\in A}\sum_{\pi\in\Pi}\sum_{\mu\in\mathbb{E}}\sum_{x\in[-2,2]\cap\mathbb{Q}} x\text{P}(\Delta V_{\mu_k}^{\pi_j}(\alpha_i) = x|\text{knowledge}_{\pi_0}(\pi_j)) \tag{2.1}$$

Where the value $x \in [-2, 2]$ of $\Delta V_\mu^{\pi_j}$ is inferred by the actor given its knowledge of the agents it can infer will be affected by its course of action, as denoted above by $\text{knowledge}_{\pi_0}(\pi)$.

With some necessary information about the actor's intelligence, an external observer π_{ext} could much more accurately estimate the likely ethicality of its decisions by considering a similar expected value (note that goals are factored into, and its knowledge of other actors is determined by, $\Upsilon(\pi_0)$), for values $x \in [-1, 1]$, and $y \in [0, 1]$:

$$\mathbf{E}(E(A)|\Upsilon(\pi_0) \approx y) = y\sum_{\alpha\in A}\sum_{\pi\in\Pi}\sum_{\mu\in\mathbb{E}}\sum_{x\in[-2,2]\cap\mathbb{Q}} x\text{P}(\Delta V_{\mu_k}^{\pi_j}(\alpha_i) = x|\Upsilon(\pi_0) \approx y) \tag{2.2}$$

If the external observer makes no direct assumption about the numerical value of the actor's intelligence, and instead acts on general knowledge of the actor, we will have the weaker statement

$$\mathbf{E}(E(A)|\text{knowledge}_{\pi_{\text{ext}}}(\Pi)) = \sum_{\alpha\in A}\sum_{\pi\in\Pi}\sum_{\mu\in\mathbb{E}}\sum_{x\in[-2,2]\cap\mathbb{Q}} x\text{P}(\Delta V_{\mu_k}^{\pi_j}(\alpha_i) = x|\text{knowledge}_{\pi_{\text{ext}}}(\pi_0, \pi_j)) \tag{2.3}$$

Please note that while ΔV_μ^π is discontinuous and countably infinite, π represents a

discrete agent and thus we are taking its action and knowledge, as well as knowledge about it, to be discrete random variables.

Furthermore, an ostensibly unaffected observer (or the actor itself) could, given the intelligence of the actor as approximately b for some $b \in \mathbb{Q} \cap [0, 1]$, give some friendly advice about the likely ethicality of a course of action by considering the following, as a function of its own intelligence. Note that the actor need not know or infer its own intelligence precisely; it may simply have some self-perception or opinion of itself. It is irrelevant whether or not this introspection or set of inferences is given, as the given intelligence can be replaced by $\Upsilon(\pi_0) \approx y, \Upsilon(\pi_i) \approx z$ with the approximation being what the actor can infer about its intelligence and that of another agent. However, the actor's intelligence or an approximation thereof is a *sine qua non* for an observer to measure the probability that its course of action will have an ethical outcome. Furthermore, it is clear that the observer will have different inferential capabilities than the actor, such that they will both obtain different probabilities, with a probability proportional to the difference in intelligence between them, that the most accurate will be the one conducted by the agent with higher intelligence. It is important that we consider self-knowledge as an aspect of intelligence, since an agent that knows its own limitations will be less likely to risk an action with lower probability of ethicality or success if it chooses to be ethical.

The importance of the relationships we have outlined between intelligence and ethicality uphold the orthogonality thesis while providing an accurate representation of intelligence applied to ethicality: The unethicality caused by a lack of intelligence will simply not be as great as that caused by high intelligence, not because the damage may be less; humanity is more likely to end by its own stupidity than by some catastrophe, but this measure ensures that a highly intelligent agent set on being unethical for whatever purpose will be a much more pressing threat than an unethical actor with low intelligence, or a well-meaning but unintelligent agent that

has produced unintended consequences. Higher intelligence, especially as it goes to superintelligence, is much more difficult to contain. The matter of reasoning it out of being unethical, or some concept of redemption is out of the scope of the current work.

2.2.1 The Environment's role in ethicality

Since V_μ^π requires an environment, then it is clear that environments should play a role in ethicality: This would hold even with utility functions; given bounded rationality and the fact that all agency is embodied (there is no such thing as an incorporeal agent), our goals and expectations are bounded by the environments we interact with. More importantly, these environments (and we could consider ourselves a reward-summable environment, when we ponder our internal states and psychological reward mechanisms) have a *causal history*. There is a degree of determinism in decision-making, which therefore applies to ethics. As an example, while stealing is always unethical because it affects the expected future discounted reward of an agent, the change in this expectation will be less if, say, one's coat is stolen in an idyllic beach as opposed to a snow-covered mountain. For example, if one steals food to stay alive, and the affected agents lose very little, the moral responsibility would be low. One could further argue that a person who needs to steal to survive has diminished agency and intelligence due to the pressing time preferences that brought them to that situation.

Notice also that V_μ^π is used in calculating intelligence, and therefore the ethicality of an action can be used to measure the intelligence of those affected by it both in how they deal with the impact, and in the mere fact that their expected future discounted rewards have changed. This means that different environments will not only change the measure of ethicality, but that our actions impact the environment and intelligence of others, giving weight to the importance of Duties of Perfect Obligation, since these are negative statements and violating them would result in a categorically unethical

course of action that will visibly impact the agents. In other words, the destruction of the environment without proper substitutes and compensation —e.g. more optimized artificial environments— must be unethical.

This implies that, in Legg's model, intelligence is not completely innate; it is situational and specialization-dependent, such that universal intelligence is measured across environments to account for this fluidity. This is why we chose this particular model over others: Ethics must also be fluid and adaptable to the particularities of a situation and to its causal history.

2.2.2 An action's impact on the environment

Our actions have an impact on the environment, both subjectively and objectively. For the subjective environment, imagine the following situation: Suppose you are in a classroom, taking a class with your peers, or perhaps teaching one. That is a definite environment, with a specified set of goals related to your course, and thus it has a particular algorithmic complexity. However, should there be an active shooter situation, the fact of a threat radically alters your goals, and makes the environment unthinkably more complex, because now we are under threat and must therefore consider the related variables, escape routes, etc. This may be a drastic example, but even something as simple as stealing someone's coat, or, on the ethical and proverbial side, teaching a man how to fish, will alter the complexity of their subjective environment. Therefore there exists a category of actions which clearly change the complexity of the environment, meaning that if we follow through with Legg's definition, for these actions we would have that (with sufficient information) the ethicality of an action can be most accurately measured as the change in intelligence of the affected agents per action. Thus, since $\Upsilon(\pi)$ already factors in the environment, we would have that

$$E(A) = \sum_{\alpha \in A} \sum_{\pi \in \Pi} \Delta\Upsilon(\pi_j|\alpha_i)\Upsilon(\pi_0) \tag{2.4}$$

This is also implied by previous definitions which do not affect the complexity of the environment: To a higher or lower degree depending on the particularities, an action will inevitably have an impact on the intelligence of other agents. Notice the largest flaw right away: It requires the calculation of two intelligence functions, and as such the expected ethicality, specifically (2.2), is much more tractable than the actual ethical outcome (2.4), as the former requires less computational power.

2.2.3 The basic components of ethicality

Finally, let us analyze ($E(A)$) given this latter definition and study how an agent can infer the ethicality of these actions. We should identify two separate terms: The term that makes up the intelligence of each π_j and the one that is the intelligence of π_0:

$$E(A) = \sum_{\alpha \in A} \sum_{\pi \in \Pi} \Delta\Upsilon(\pi_j|\alpha_i)\Upsilon(\pi_0) \tag{2.5}$$

$$= \sum_{i=1}^{|A|} \sum_{j=1}^{|\Pi|} \left[\sum_{k=1}^{|\mathbb{E}|} 2^{-\Delta K(\mu_k|\alpha_i)} \mathbf{E}\left(\sum_{l=1}^{\infty} \Delta r_l^{\pi_j} \right) \right] \left[\sum_{k=1}^{|\mathbb{E}|} 2^{-\Delta K(\mu_k|\alpha_i)} \mathbf{E}\left(\sum_{m=1}^{\infty} r_m^{\pi_0} \right) \right] \tag{2.6}$$

$$= \sum_{i=1}^{|A|} \sum_{j=1}^{|\Pi|} \left(\sum_{k=1}^{|\mathbb{E}|} 2^{-\Delta K(\mu_k|\alpha_i)} \right)^2 \left(\sum_{k=1}^{|\mathbb{E}|} \mathbf{E}\left(\sum_{l=1}^{\infty} \Delta r_l^{\pi_j} \right) \mathbf{E}\left(\sum_{m=1}^{\infty} r_m^{\pi_0} \right) \right) \tag{2.7}$$

Here we have several probability distributions, namely those of the rewards $\forall \pi \in \Pi$, and yet one more if we account for the possible changes to the environment, though the former distributions are subject to the environment itself. The issue is that we have no guarantee that the goals of these agents are independent, and the expected value will depend on how the random variables behave. In order to approximate these expected values during an emergency situation, a machine would most likely attempt to estimate from a prior distribution, or approximate given the data gathered by other agents in similar situations.

2.3 What do we do about solipsistic AI?

What if an AI is not aware of other agents? This is the case with contemporary AI, and the lack of awareness of other agents is called solipsism. In philosophy, this term does not signify lack of awareness but rather the idea that the only conscious subject is oneself, but even if the AI is not conscious by some philosophical standards, its inability to measure the impact of its actions on other subjects is what makes it solipsistic and dangerous.

First of all, what about agents that cannot make their own decisions, such as children, the mentally ill, or the mentally deteriorated elderly? Well, these are supposed to have a guardian precisely for that reason: They can harm themselves or another unwittingly, but do not have the intelligence or agency to accurately assess the consequences of their actions, which means it is not only permissible but morally preferable to have them under the stewardship of a guardian.

These examples help when thinking about solipsistic AI: Even if the examples are aware of other agents, they are unaware of consequences and are not morally responsible for their own actions. Furthermore, with or without decisive strategic advantage, solipsistic AI is dangerous due to its lack of awareness of other agents. The problem is that without that awareness, an actor cannot calculate the expected future discounted rewards or intelligence function of the agents its action will impact. Even if we were to hard-code an ethical algorithm, it wouldn't be able to infer its impact on others. Therefore, any such AI that is sufficiently powerful or dangerous must be overseen by a non-solipsistic decision maker which can calculate the ethicality of its proposed course of action. Since the AI is solipsistic, it is debatable whether or not it would be able to fool the overseer, but it is clear that these AIs need to be monitored carefully.

The problem is that there are powerful and dangerous narrow AIs (meaning AIs that are dedicated to a single task) in existence already, such as autonomous vehi-

cles, which are known to need supervision —be it hard-coded and pre-decided by programmers, or user-defined.

Of course, not all solipsistic AIs are a danger: A contained AI such as IBM's Watson supercomputer, or at least the AIs contained therein for, for example, natural language processing, is fairly harmless. Again, the only risk here is that these machines are not responsible for their own actions and so we must interpret their actions and direct the machines with extreme care.

That being said, an external decision-maker is not a good answer: For one, an external agent will have its own biases and risks. Furthermore, given that understanding does not have measurable effects on outcomes, an AGI does not need to understand that we are alive, what that means, what intelligence means. It merely needs to identify us in some form (e.g. through its sensors and some algorithm for classification), and have a way to estimate the potential benefit or damage it might do to us in the pursuit of its goals.

2.4 Example: Self-driving cars

Such a system most likely does not have the time to estimate pedestrians' intelligence when it finds itself in an ethical problem. Instead, it is easier to calculate different ethical outcomes by calculating the likely change in their overall ability to get what they want given the probability of survival, or the likely extent of the damage to any given pedestrian (and the driver) given the rate of deceleration of the vehicle. This works best if all self-driving vehicles were connected and constantly sharing information with a main database that updates the probabilities of survival as it gets more data from accidents. But we must address the buzzword: Where does responsibility lie when a self-driving car or any other automated system is involved in a death?

2.4.1 Moral Responsibility

When dealing with this same issue, Tigard (2021) defines three kinds of responsibility: Normative, which is behaving according to socially accepted norms, possessive, which refers to duties and obligations, and descriptive, which is "describing in detail the potential status of the subject, namely as the source or cause of something that happened." Another important distinction is that between being responsible and being held responsible, with the former being inwardly posited, and the latter being externally imposed.

A simple assumption also follows: An agent cannot be morally responsible in any way shape or form for consequences it cannot infer, or the outcomes of goals it cannot choose. This comes with the caveat that we are not making statements about other categories of judgment; our arguments and assumptions have no bearing on, for example, whether or not there is a need to punish those who are not responsible for harmful acts they have committed, or to isolate them from society. Likewise, nothing in our work will provide a way to deal with those who behave unethically.

In other words, while we can share a large set of moral intuitions, as well as interpersonal, cultural, and institutional agreements, all these things are not formalizable and a properly universailzable ethics may have to forego our intuitions. Cases such as the feminist bank teller fallacy (conjunction fallacy, in which a set of conclusions is taken to be likelier than a single conclusion) show just how flawed and unreliable our intuitions are, and how bad humans can be at mental statistics; the case of computable morality is akin to that of universal intelligence, because the latter 'does not agree with some everyday intuitions about the nature of intelligence' (Legg, 91). In order to develop such an ethics, we must immediately note that all we have said thus far means that an ethical decision is fully dependent on the inferential capabilities of the agent. That is to say, the ethical process depends on inference, and will necessarily come with a degree of uncertainty, minimized as intelligence rises. The

outcome of any given decision can be characterized as the expected value of another agent's (or agents') reward function after a decision given the intelligence of the actor making the decision. Such an ethics would be falsifiable and universailzable, while at the same time being specific enough to each agent's conditions by contemplating its capabilities.

There are many factors that determine how we infer the reward function of other agents, from straightforward communication to whatever mental models we build about how individual x reacts to situation y based on the aggregate of our experience with said individual, then generalized to how people in general react to possible situations. In other words, we as humans are not able to calculate the measure, only to approximate it as 'better' or 'worse' than the point of reference (inaction), and roughly how much better or worse comparatively. The importance of dialogue comes into play right here, because, as we will later argue, the way to build a better model that can more accurately approximate the reward function is in dialogue with the possibly impacted agents, or other agents. The more intersubjective experience we have, the more we can learn from each other, and the more material we have to simulate new scenarios. Both these methods are useful adaptations that help maximize cooperation and thus our utility functions, but they all rely on a deeper factor: Intelligence. How much we can infer from another agent is limited by the accuracy of those inferential abilities, and since this inference is central to achieving our goals, it is in reciprocal presupposition with the intelligence function.

Ethical theories are supposed to provide guidance, yet it seems difficult to follow ethical theories on their own, much less take them seriously, if they have no factual evidence or computation that we can use to verify that these decision-making systems are maximizing our benefit. In other words, to generalize away from the relatively prevalent moral intuitions of humans, we must think of an ethics that is computable and based on possibilities for decision-making.

In short, the moral responsibility of an actor *regardless of whether or not it chose to act ethically* is an approximate value to the actor, but more accurately verifiable for an external observer. It must be equal to the expected ethicality of the action given by the intelligence of the agent —so long as $\Upsilon(\pi_0)$ is given. Notice that this holds for an agent who chooses to be either ethical or unethical; an agent that chooses to be neither has simply chosen not to act.

On the other hand, while the "criminally insane" are not liable because they are assumed unable to understand the context of a trial, or the legality of their actions, from an ethical perspective even this consideration is irrelevant: We may seek a way to treat such individuals or seclude them from society, but fundamentally the extreme nature of their case is such that they do not meet the conditions to be considered moral agents —yet they could potentially become moral actors under the right conditions, such as medication and other forms of treatment.

Also notice that, in a way, this equation factors in an *ought*: That is to say what an agent of a certain intelligence *ought* to be able to infer. This, however, is not a traditional, normative *ought*, since it is dependent on the actor itself. It is not enough to infer consequences without using one's intelligence to the best of one's ability, given the situation and environment. That is, if being ethical is a goal, or the ethicality function is used as a reward function, then to go about it most efficiently an agent must dedicate as much resources to the decision as possible. In human terms, this would be equivalent to the Kantian imperfect duty of self-improvement in humans. The more *fit* we are, physically and mentally, the more we are able to aid others and act more ethically. Kant's arguments is that we rely on experts all the time, so we ourselves should strive to be a reliable expert if we wish to act in a universalizable fashion.

2.4.2 Are Machines responsible?

In spite of the previous discussion, we must ask ourselves this question more carefully. We must wonder whether machines assert anything at all, and if they are able to posit an ethic or a theory of truth. Take a Large Language Models such as GPT: Given a prompt, the model outputs the answer that has the highest probability of being a correct sentence in the language it models given the prompt; it is not endorsing, asserting, or believing in what it says. There is no will, good or bad.

This means it makes no sense to hold them responsible; there are no means for justice, let alone one of the restorative kind. Prevention through creating the best possible ethical algorithm we can is one of the ways to address the issue.

Chapter 3. Existing human moral systems: Explanations and comparisons.

In this chapter, we will evaluate ethical systems by different standards: The human, and the scientific. We therefore ask the reader to not take anything as the final word for, as our exploration of each system will show, all the systems explored here (with the exception of those that are examples of the value-loading problem) have their weaknesses and strengths: Qualities that we will criticize at the human level may be redeemed at the analytical level, and vice-versa.

We have to get Friendly AI right from the start. That much is clear: Should an unfriendly AGI achieve decisive strategic advantage, come about first, then that will be it. Luckily, this discussion has become more commonplace in the past twenty years, to the point where it would almost be redundant or alarmist to hammer on the topic right now. By itself, however, even friendly AI is not enough. We will still have to keep up. Not only do we have to rethink our intelligence and professions to keep humanity productive —not everyone will be content with the blissful idleness of the ideal singularity, and there is no Marxist Eden of self-realization—we will also have to completely reframe the pragmatic variables of our existence.

The problem with enumerating our values to an AI has been explored before, but bears restating. As explored by Yudkowsky (2011), it is twofold: Firstly, by the problem of induction we can never be sure that we have properly formalized each and every single value, and a similar problem is in play for meta-values. More importantly, our ineptness at specifying what we value —and from a psychoanalytic perspective the self-frustrating nature of our desire, born from lack rather than aiming at an existent object— should alert us to the idea that our values themselves are flawed or at the very least lacking and optimizable. If we are to achieve some form of indirect normativity, some "process for deriving a standard" (Bostrom, 2014, 163) with which to judge values and interactions, then we cannot be overly fond of the ones we currently uphold; they must be questioned and reevaluated according to

more stringent formalized standards —this indirect normativity is what could take the place occupied by our current legal systems, for example, and once again the Kantian Categorical Imperative is the most air-tight indirect normativity we know of. Thus, the question is just as much "What do we want an AGI to do for us?" as it is "What do we want ourselves to become?" The state risk accrued by our current insufficient values will soon catch up to us, if Nietzsche's work on the reevaluation of values and the endless complaints raised by idealists who wish to 'build a better world' through some political movement are anything to go by.

Ideas such as justice, compassion, equality, happiness, empathy, and authenticity have no place when thinking about a future that is post-cultural, in the sense that the anthropocentric aspects of our current cultures will not hold anymore. Happiness will always have a place for many, but it is not an argument, and it is risky to posit it as a value (The obvious hyperbolic example: Mandatory heroin injections). As soon as a nonhuman element is introduced, these assumptions become dangers, weak links fostering miscommunication at best and existential risk at worst. The problem is that these 'values' are not values in themselves, in the sense that we cannot argue for them from either a deontological or consequentialist perspective because they cannot be properly formulated: On the one hand, different people hold diverse and contrary conceptions of each, all justifiable under the current status quo, and on the other we have so many values that they are bound to clash with each other, lest we find redundancies and reduce some values as instances of others until we find a stable meta-value.

Theologians such as Boethius would argue that the problem is much simpler: Earthly values and pleasures are many (Wealth, beauty, power, and so on), but the perfect joy is one, and all these values could be seen as attributes or manifestations of an underlying oneness that beckons to us from some metaphysical space. A monotheistic approach could well be seen as a crude attempt at indirect normativity: Even

if an institutionalized religion lies down 'laws' to be followed for safe passage into a happy afterlife, these norms are often abstract forms that signify ways of being in the world that ensure the better functioning of a community —as proof, consider the notion that a priest spreads his genes indirectly, by shaping his community, and this is a legitimate evolutionary/game-theoretic strategy. Sadly, we no longer have the luxury of waving our hand and appealing to an afterlife or some deity: Even if it were to exist, we cannot defer or attribute our values onto it because this will not solve the problem of creating a formal structure for a friendly AGI.

So, for now, let us focus on ethical systems that are meant to be internal ethical generating functions within the agent. These generally admit intersubjective communication; given the asymmetry of intelligence and its orthogonality with goal-orientation, a Kantian may hold that it is a moral imperative to communicate our goals and values with others who may find logical inconsistencies we have not detected.

3.1 Intrasubjective Ethics.

A possible resource we might be able to offer is to compare several ethical theories, to see which of these manages to contain and expand upon others in a manner that is to some degree amenable to being falsified, with the assumption that if that theory fails, then all those it contains must fail as well, as particular instances of the theory that contains them. Having done this, the next chapter will attempt to formalize the strengths of pre-existing systems, to see what strengths and weaknesses our proposal has, and modify it accordingly.

It is the general tendency of scientific theories to both fully contain and expand on previous theories, such that all the phenomena previously explained are explained more accurately and adding to that the ability to account for phenomena which previous theories could not. There is an obvious difficulty in translating these criteria into a philosophical context, insofar as philosophy does not necessarily need to account

for falsifiability, and as such it does not necessarily follow that a theory is better simply because it accounts for previous theories; the criteria for proof and validity are often different from those of science. Nevertheless, if a theory can account for previous ones and adds more stringent methods of proof for its validity, it follows from this account that it is at the very least preferable to the previous ones.

3.1.1 Aristotle, Kant, and Mill

In this section, we aim show that that Mill's utilitarianism fully includes Kant's deontological ethics and Aristotle's virtue ethics, as well as adding an element of falsifiability and providing a means of mathematical analysis, which makes his theory preferable to Kant's, even if it is not necessarily optimal or applicable when formal considerations come into play, as we will consider later on. However, as our proposal is fundamentally a calculation, if Mill's system includes and expands on Kant's and Aristotle's, then so does our proposed measure.

Kant argues that freedom is found in constraining one's actions to morality in accordance with reason and the principle of non-contradiction (Critique of Pure Reason, 388): He defines duty as "the necessity of an act done out of respect for the law." (AK 4:400) From this he defines the law as the Categorical Imperative: to act as if one's actions were universailzable.

> "To do no action on any other maxim than one such that it would be consistent with it to be a universal law, and hence to act only so that the will could regard itself as at the same time giving universal law through its maxim." (4:434)

This follows from his definition of reason as the "faculty of the unity of the rules of understanding under principles." (B359) This allows Kant to claim that "the true vocation of reason must be to produce a will that is good, not perhaps as a means to other purposes, but good in itself for which reason was absolutely necessary."(4:396)

Thus, to Kant, moral action is an end in itself, and thus its worth is absolute:

"The moral worth of an action done out of duty has its moral worth not in the objective to be reached by that action, but in the maxim in accordance with which the action is decided upon; it depends (...) but solely on the principle of *volition in accordance* with which the action was done, without any regard for objects of the faculty of desire." (4:399-400)

Furthermore, he characterizes empirical considerations as infringing "upon the purity of morals themselves and [proceeding] contrary to its own end." (4:390)

This means that, though logical, Kantian morality claims to be incompatible with practical concerns and, at face value, resists assimilation into other theories. And yet an empirical claim can make a case against Kant: We could accuse him of certain optimism, insofar as he seems to claim that all humans are sufficiently intelligent to analyze their behaviour with such rigour and act accordingly. Furthermore, as Hegel claims, one is not a philosopher merely because one has the ability to reason, and the fact that one can reason means little or nothing without rigorous training, and not everyone can question their deeply held beliefs, contradictory as they may be (PS §66). Mill accurately identifies this problem from another perspective, that of trauma —though he does not call it that way— which can stunt reason and intellectual growth, or turn one away from morality and public utility by creating the impression that it is contrary to one's own benefit, thus making people go for the 'nearer good.' (Utilitarianism, p.13) This is supported by research from all fields of psychology, especially clinical psychology, which have directly proven the impact of childhood trauma on brain development, and the fact that IQ has an upper and lower bound through genetics —though there is a correlation between higher IQ and higher rates of depression, a satisfactory answer to this correlation has yet to be found— and a good education and raising environment can help reach that higher bound, which is something Mill —with the tools available to him at the time— also identifies, providing us with a partial solution through the ideas of habit and education. (14,

16)

We posit that it follows that Mill's theory includes and expands on Kant precisely by including these empirical considerations which help address this disparity in intelligence and training, such that all men may be able to be more moral even by Kantian standards, through the improvement of the human condition by maximizing pleasure and minimizing pain, while noting that it is impossible to constantly maintain a state of pure bliss, just like it is impossible for embodied beings with heterogeneous impulses and degrees of pain and pleasure to reach a state of pure rationality. (14)

Mill attempts to assimilate Kant's theory by addressing the empirical dimension of rational pursuits through "their circumstantial advantages rather than in their intrinsic nature." (11) There need not be an inherent value or nobility in the pursuit, which is impossible to prove —the claim that anything has intrinsic value is an extraordinary claim, requiring extraordinary evidence, since we are constantly assigning and negotiating value subjectively: It is simply that these human faculties are less susceptible to material need than more 'animal' concerns. Thus Mill appeals to man's sense of dignity; Kant defines this dignity, in accordance to reason, as proper to "a rational being who obeys no law other than one which he himself at the same time gives" (4:434). This definition is useful to explore how Mill's theory includes Kant's, insofar as the preservation of this dignity, granted by individuality and 'reason,' is an important element for the state of happiness, and thus the furthering of reason is a utilitarian goal, such that acting in accordance to the categorical imperative is fully compatible with utilitarianism, if only due to it being conducive to public utility rather than for its own sake. In Mill's utility, we find that happiness is increased gradually, and thus people would become more moral by Kant's standard over the course of time.

Aristotelian virtue ethics are rather simple to characterize, and Mill subsumes them easily through public utility. Whether virtue is god-sent or learnt is irrelevant to

Aristotle's argument (Nichomachean Ethics, 1.9). The central idea of virtue ethics is developing one's character and attributes (virtues) in order to have the best possible conduct. This is achieved by both controlling one's passions and cultivating one's mental capacities, and it all works to aim towards a lasting, transcendental happiness (finality), since all other things are, to Aristotle, a means to an end. This is what makes it easily assimilated by utilitarianism: "If we are told that its end is not happiness, but virtue, which is better than happiness, I ask, would the sacrifice be made if the hero or martyr did not believe that it would earn for others immunity from similar sacrifices?" (18) That is to say, virtue in the Aristotelian sense is a proper subset of Mill's public utility function, and from Mill's perspective it can be argued that the private utility function has several intersections with the public, one of those being the cultivation of these virtues.

While Aristotelian ethics must be discarded due to being a clear example of value enumeration, this ethical system gets something very important right: Developing one's characters or attributes, also known as *cultivating* oneself, is fundamental to what the Greeks would call "living a good life." In more unassuming terms, we can say that self-improvement nurtures intelligence and enhances agency.

Now, as we have hinted at earlier, Mill's inclusion of these theories is not sufficient to prove his own theory superior. Kant's framework resists assimilation for the reasons we have pointed out, and Kant would most likely not be satisfied because reason and morality would not be held as ends in themselves. This is where falsifiability comes to the fore as an element that, though not sufficient by itself, helps strengthen the case for Mill.

Given that Mill grounds his theory on a fundamentally historical and material claim, his theory is vindicated by the fact that social and economic conditions have been steadily improving since the industrial revolution, and clinical psychology has shown that, after securing an income that covers one's needs and leaves room for con-

tingencies, higher earning does not correlate to higher self-reported life satisfaction or happiness, which shows that being able to satisfy one's needs is fundamental to happiness itself and all other pursuits —it is quite hard to reason on an empty stomach. While the earning gap has widened, and higher relative poverty is correlated to higher crime rates, nobody would want to return to the living standards before the industrial revolution; the lower classes today have better living arrangements than even the pre-industrial nobility.

Though Mill's theory can be mathematically tested, there is a case to be made that the system of happiness is simultaneously partially deterministic (due not only to the reasons we have outlined here but also to bounded rationality and other such concepts) and unpredictable.

Another consideration we must keep in mind is that the Kolmogorov complexity of any given consciousness is that consciousness itself. This means that it is impossible to compute the complexity of any given mind, and lends credence to Mill's argument that pain and pleasure are heterogeneous throughout the distribution of consciousnesses, as well as being a scientific account for diversity in sources of happiness and different reactions to any given condition —in short, it is strong evidence for qualia— and therefore another argument against Kant's normativity.

The evidence makes it likely sufficient to prove that Mill's theory is falsifiable. However, none of this necessarily proves it to be correct. For one, just like IQ research, they are evidence that there can be no truly utilitarian public policy, no one size fits all solution, both because it is impossible to account for the consequences and because the individual paradigm is too complex to assimilate into practical collective action. They are, however, a strong argument for Mill's theory in issues such as education, for they help make the case that if education becomes more individualized there can be an improvement in happiness through the improvement of the individual mind, aiming towards a higher rationality and thus a higher morality by Kantian

standards as well. Moreover, the steady improvement of human living conditions is exponential, following the pattern of scientific progress and according to Kurzweil's law of accelerating returns, but neither are a product of intentionality insofar as science is not intentional or teleological: One can rarely predict what the next invention will be or when certain barriers will —or even if they can— be broken, much less what the consequences of such an invention will be, and the AI and robotics literature are a perfect illustration of this. Public policy, on the contrary, is an intentional process; an attempt to control without fully understanding. Even the useful projects like nuclear technologies and NASA have the state as an investor, and make no promises on findings, operating like publicly funded enterprises rather than acts of public policy. A good example of what we refer to as public policy would be education, a consistent failure whenever generalized approaches are taken. Our argument is the same as that of many IQ researchers: We are against such public policy that would enforce such a thing as a one size fits all solution.

Making the above point is important because it shows that though we have proven Mill's theory to include Kant's, account for things it cannot, and provide additional methods of testing, this does not mean his theory is directly actionable on a public level or necessarily correct —nor does it show that previous theories were incorrect; to assume that would be 'wronger than wrong,' except for theories grounded in speculation about the transcendental, or metaphysical claims which are, by scientific standards, not even wrong. Just like the introduction of a new scientific theory, the additional methods of proof and testing point the way towards a larger set of problems that must be faced, as well as the necessity to formulate new theories that may help us deal with these problems; this is why science and technology are valid ways to approach ethics, because their progress is in this case analogous and technological progress is correlated to a higher degree of happiness, but also —as Marx pointed out— new ways to interact with our material conditions create new needs. This also

proves Mill's claim that an absolute state of happiness is impossible: We may objectively be better off, but progress is neither linear nor teleological; the introduction of new ways to satisfy our needs brings with it the creation of new needs, new problems to address.

Do note, however, that from an act-utilitarian perspective indirect normativity that focuses on not impacting other agents' utility functions, or respecting negative rights, would make surveillance not only morally permissible, but also desirable. The issue is that humans may believe to have sufficient schemata with which to infer the utility function of another agent —meaning that we can sufficiently project a set of assumptions about what may or may not negatively impact another agent's utility function— but non-intrusive mass surveillance would provide more data with which to infer other agents' utility functions in order to take decisions which are most conducive to one's goals without negatively affecting them. Even so, the problem with enumerating values shows that no amount of surveillance would be sufficient to truly infer another agent's utility function —assumed to include its value system— due to the problem of induction but, given a game-theoretical framework, gathering as much data as possible seems desirable to maximize utility and ethics. Surveillance would facilitate reaching a maximally ethical solution, defined as a set of actions which maximize utility while minimizing the negative impact on another agent's utility —since absolute prescience is impossible we assume that whatever an agent is unable to predict with its current intelligence is outside the set of ethical decision-making factors, such that absolute ethics or utility are impossible. The question is whether the gathering of psychographic data (e.g. Facebook-Cambridge Analytica data scandal) can be considered to provide sufficient information to a superintelligent agent.

3.1.2 Two falsifiable theories: Rawls and Mill.

If we assume that an ethical theory could be decided as better than another one if it contained and expanded on the previous one while adding an element of falsifiability

and the possibility for mathematical analysis, we are confronted with a series of problems: Any empirical dataset is insufficient as direct proof —outside of abstract mathematics, we have the problem of underdetermination by evidence— such that when confronted with another theory which also contains and expands on previous theories and adds a mathematical and material element, we need a different approach to decide whether one can possibly be better than the other. Testing instantiations of each theory can provide evidence of the better functioning of one or the other but this is insufficient so long as the underlying mathematical system of either one is proven wrong, a wider theory is posited, or as is the case in this example, we analyze the sets of underlying assumptions.

This, however, is a tool of mathematicians, and we are essentially pitting two philosophers together, so the question becomes whether, since it is not within our immediate possibilities to make a proper mathematical analysis, there can be any standard of proof such that we could determine one theory as better than the other. With this in mind, the avenue we will explore in this section is whether or not these two theories share presuppositions, whether their assumptions are falsifiable or axiomatic, and whether these axioms are justifiable.

Rawls's fundamental mode of arbitration is having oneself as a mediator, such that the good can be decided from behind a 'veil of ignorance,' which he terms the original position (A Theory of Justice, 11), and which means that one must bracket one's privileges and wants so as to decide social principles and norms without bias. This, however, is not precisely possible; just like with the phenomenological epoché, one cannot bracket language, one cannot bracket one's history insofar as one is in part the product of personal history which is a product of historical and material conditions. This mode of arbitration is, to us, flawed in that regard.

Rawls argues that utilitarianism would sacrifice a vulnerable minority to the greater good, and it seems to us that Rawls isn't characterizing utilitarianism cor-

rectly; he seems to go over the fact that Mill argued that few are in a place to truly do something of public utility, and most generally strive for private utility, something Rawls attributes to social justice, defined as "the principle of rational prudence applied to an aggregative conception of the welfare of the group" (21). Utilitarianism does not seem to ask for any sacrifices; Mill argued that martyrs, though beneficial and welcome if they do so willingly, are by no means a necessity, and furthermore posited that the just and the expedient are not in opposition —which might be reason, under a directive fashioned after Mill, for a superintelligence to provide us with a steady dosage of heroin. This, however, does not undo Rawls's theory, and for all of Mill's advantages in our eyes, we will now attempt to reject both:

Both Rawls and Mill appeal to the rationality of their readers, which seems rather more like an appeal to vanity to the effect that 'if the reader is rational, it will agree with me.' Their presuppositions, however, are not necessarily rational: They are both essentially social thinkers, in the sense that Rawls is a contractualist and Mill looks for public utility in the long term, through a form of gradualism —both for individuals through education and habit, and societal progress. They both argue that the greater good is found within society, striving for the good of this phantasmatic society, but in doing so they envision a certain social model. The problem here is that, though we always-already think as socialized human beings, this does not necessarily mean gregariousness is rational. The "social feelings of mankind" Mill directly appeals to are an evolutionary matter, and it is well known that one cannot derive an ought from an is: Just because humans are social, and contemporary society provides us with comforts and commodities that make life easier, does not mean that we necessarily ought to wish for higher social good or cohesion, not because they are a bad thing, but because one cannot derive an ought from an is, and oughts can carry risky assumptions with them.

This assumption Rawls and Mill share is characterized by the idea that society is

akin to a big family tied by allegedly rational bonds, that ought to stay together due to arbitrary and tribal in-group preferences —and though these preferences most likely arose due to selection mechanisms, these are not admitted, but rather operate while being disavowed. No reason is given as to why we should care about our neighbour, or even whether it is truly in our self-interest and, though both speak of individual freedoms and benefit, this is ultimately subordinated to the society itself:

> "Every community has the propensity, stronger or weaker according to the fullness of its power, to become an authority to its members and to set limits for them: it asks, and must ask, for a "subject's limited understanding"; it asks that those who belong to it be subjected to it, be its "subjects"; it exists only by subjection. (...) The society demands that those who belong to it shall not go beyond it and exalt themselves, but remain "within the bounds of legality," e. g., allow themselves only so much as the society and its law allow them." (Stirner, op cit, 153)

While Mill at least admits that few are truly in the position to serve public utility, and private utility is common and desirable, the very distinction seems to us to assume that we must all be striving for society; Rawls's original position seems to demand that we make concessions in the name of an unexamined fairness, and Mill openly argues for educating people in a way such that they will identify the feelings of others with their good, to strengthen social ties and invest oneself in society. As Mill well points out, however, these feelings are not innate but acquired, and so moral feelings are a random outcome of evolution, if evolutionary psychology is to be believed, which does not mean the feelings are proven to be a legitimate basis for a rational argument but merely something that was 'true enough' for a time in our history.

The problem with Mill's dignity, Rawls's fairness and equality, and other such concepts is that they are essentially feelings: These ideas can be posited intrinsically, but only attributed extrinsically, never truly proven outside of a subjective frame of reference. We must either project our own sense of dignity onto others, or believe

that they have it, if these ill-defined concepts even make sense to us. Likewise, there is no unmediated mediator such that fairness can be objectively decided in any given situation. As mentioned before, it is truly impossible as always-already socialized and linguistic beings to truly place ourselves behind a veil of ignorance and decide impartially. These concepts, just as consciousness, and other speculative ideas we posit intrinsically about ourselves, can only be believed to be had by others and, as posited through Bayesian probability, partial rational beliefs are essentially probabilistic. At least we are kind enough to assign partial rationality to these beliefs, though they might as well have none at all; they cannot be projected onto the whole world, and their time may well be up.

We do not posit that others do not have feelings, we are far from solipsists, but the point of our argument is that we cannot fundamentally prove their existence even in ourselves, and thus that we cannot prove that it is an analogous experience for others, much less something we can confidently transmit to an AGI in the hopes that this will yield a happy result. Simply put, given the problem of trust, it is risky to believe than an AGI will take our feelings as a matter of trust, and realize their relation to our internal reward systems.

Social cooperation works by degrees as a fluid assemblage, which should not be imposed, and rather than deciding rights and duties immediately and jointly as Rawls proposes, these can be carefully and peacefully negotiated through time, with respect for individual freedoms so long as nobody is harmed, preserving the right to be left alone. That is, if we can build an AGI that values patience and negotiation, which seems to go counter to the speed of RSI. Incidentally, the right to be left alone, to be allowed not to interact, could well be a fundamental safeguard against a utility maximizing superintelligence seeking to drug us all into bliss. Imagine, for example, an AI that asked for our consent before making a decision that affected us!

Most importantly, perhaps is that the "veil of ignorance" implies bracketing even

our intelligence in favour of the above described feelings, which makes it highly risky when dealing with non-human agents; it is the most human of the theories described here for that reason, and therefore it needs to be discarded.

3.1.3 A practical example of intrasubjective moral systems

"Consider the following situation: As part of your quality-control job, you have been asked to study the behavior of data entry clerks at a large company. All clerks have been informed of the study and have been promised anonymity as part of the study process. During your study, you have (with the consent of management) installed key-logging software to track behavior during transaction processing using a new software package. Your analysis of the keylogs has shown that about 10% of the clerks are making a particular error; this error is both costly (to the company) and could expose the company to legal liability. Company policy also states that clerks' pay can be docked for this error. When you report the preliminary findings to your supervisor, she demands that you turn over the log files so that the specific employees can be re-trained (disciplined if necessary) and to ensure the company is reimbursed for the lost profits." (Personal communication from Dr. Patrick Juola, 10/07/2020)

In principle, no (properly interpreted) workable ethical framework can justify turning over the log files to one's supervisor. This is due to a need for some consistency across workable ethical systems and common moral intuition; most ethical theories tend to provide guidance that is congruent for most cases, but only differ in edge cases. We will show that no ethical system can justify, let alone mandate, taking over the files.

First, an example that would make turning over the files an imperative: Legalism. Given that one is an employee, bound by a contract, one must hand the files over to my boss unless it is in violation of one's contract or a 'higher' (state/federal) law. The central problem with legalism has already been discussed: The law is fickle and changeable, such that it is not a properly intrasubjective framework. It's on us to

show that no proper ethical system will agree. We must remember that ethics is, in a way, a form of searching for better ways to live together. That's why treating other people as ends rather than means is so important to Kant, and so we will start with a Kantian analysis of the situation; in a Kantian framework, there are two key problems that make turning over the log files immoral:

(1) The employees were promised anonymity; breaking a promise is not universailzable as an ethical principle. Breaking promises is explicitly discussed in Kant's Groundwork for the Metaphysics of Morals [AK4:401].

(2) While there may be room for debate as to whether or not any company is ethical, and if there can be capitalism without treating other people as means to an end rather than as ends in themselves [4:428], this situation clearly consists on using other people as means. The clerks are not acting in bad faith, but rather making a mistake; to turn over the logs would be to treat them as means rather than ends: It holds them responsible for a failing that is not their fault, treating them as means for the company, which may even discipline or garnish their wages without any intentional wrongdoing on their part.

The moral choice from a Kantian perspective would thus be to re-train everyone, or avoid disciplining and seeking compensation in order to find a solution that would not violate the promised anonymity while still spending only the necessary amount of money on retraining rather than retraining everyone.

These Kantian points actually hold for other theories, even if they approach it from a different perspective, but it is precisely because they arrive at the same conclusion through different means that it is useful to hold each ethical system as a tool; just saying that something is 'wrong' or 'immoral' is not enough.

The simplest case is that of virtue ethics, wherein lying/breaking promises is blatantly unvirtuous. To Aristotle, there would be a need for a proper balance of truth-

fulness and justice, where justice would be to find a way to deal with the employees without being too strict and vicious; the same solution as the Kantian perspective. However, note that we have discarded this form of ethics as it falls for the value-loading problem.

In yet another example, act utilitarianism would put the moral issue in terms of a calculation: The reimbursement the company can get has to be put in balance with the cost of preserving anonymity and retraining. However, there is a greater cost to turning over the log files, both for oneself personally and for the company. Let us assume (without loss of generality) that the clerks know that we were conducting the study personally:

Anonymity was promised to all clerks, and 10% of them is by no means an insignificant number, especially in a large company. What will happen to the work environment when that promise is broken and 10% of the clerks are punished without having done anything wrong intentionally? This will degrade trust in the company, which will negatively impact public utility as other departments see the actions the company took, and this in turn will negatively affect the work environment and company culture, potentially leading to losses both in terms of lost earnings (assuming happy workers, or at least workers who can trust their company, work better), and also potentially in terms of personnel who might choose to leave and go work for people who treat their employees better.

On a personal level, and this works for egoist ethics as well, if we have personally promised anonymity, and it is known that the boss was informed about who was making the error, then this will negatively impact our social standing within the company, thus directly diminishing our utility function. Game-theoretically, handing over the files is the wrong thing to do.

This consequentialist approach allows us to go deeper on the Kantian/deontological approach as well: There's nothing wrong with saying that "you should act in the best

interests of the company that employs you" is universailzable; from a consequentialist perspective, if the company does well then so do we. The utilitarian framework helps us prove that this imperative is incompatible with *always* doing what the boss tells us, which is not universailzable because the boss can be wrong about what is best for the company, and the boss might wish to do something immoral. Since the company's utility function is an emergent property of that of its employees, that which is best for the company must include the welfare of said employees, and the utilitarian approach helps us prove that we would, as a matter of fact, be acting in the best interests of the company by not turning in the log files, or by negotiating an ethical way to use this information with my boss.

Social contractualism leads to a similar conclusion, especially since the idea of breaking promises can be achieved from behind the Rawlsian veil of ignorance, and it is blatantly unjust to punish people who did not even know they were doing something wrong; it is the company's responsibility to train its employees properly.

These three examples of major theories help illustrate the broader point: Ethical theories are interconnected; they are tools to approach a situation from several perspectives, and good theories can strengthen each other's arguments. In this particular example, any system that rejects blind obedience of authority should be against turning over the log files. From a deontological perspective, it is enough to say that breaking promises and treating other people as means rather than ends is morally wrong. From a consequentialist perspective, we must remember that the calculation of potential consequences is not merely economic but also social. Therefore, a draconian attitude such as disciplining the 10% of clerks that made a mistake and forcing them to reimburse the company is more likely to negatively impact the utility function of the company as a whole by creating discontent in the individual workers. Any ethical theory that demands we comply with our boss would implicitly be forcing us to disregard the best interests of the company, boss, and fellow employees.

While game-theoretic systems requiring some sort of calculation will tend to include other workable ethical systems, having more ethical and philosophical tools available to us will be conducive to conducting a deeper and more nuanced analysis of any given situation. This helps us build a form of indirect normativity based on literacy, which will be central in the following section as we discuss intersubjective ethics.

3.2 Intersubjective ethics, Habermas' Discourse Ethics

An expansion on why rights and fairness cannot be moral arguments and a transition to applying the AGI problems to evaluate ethical systems

Habermas' enterprise starts from a common point; no proposition is self-evident, and "there are no indubitable "starting points" beyond the bounds of language, no experiences that can be taken for granted within the bounds of reasons." (Truth and Justification, 36) In other words, the long-established nonexistence of the epistemological given and the problem of language laid out by the likes of Wittgenstein, Nietzsche, and most importantly Wilfrid Sellars.

Prima facie, the idea of basing an ethics on debate with each other sounds quite attractive: Doing so allows us to make further inferences about the preferences and expected future discounted rewards of other agents. Sadly, nothing is ever that simple.

He presents his standards for deriving a truth (p.37, the italics are ours) as "rigorous (. . .) based on the idealizing presuppositions

(a) of public debate and *complete inclusion of all those affected*;

(b) of equal distribution of the *right to communicate*;

(c) of a nonviolent context in which only the unforced force of the better argument holds sway;

(d) of the sincerity of how all those affected express themselves."

(e) (Now our letter notation) It is of supreme important for Habermas that "those participating [decenter] their cognitive perspectives." (38)

That is to say, put themselves aside as a function of the pursuit of knowledge. Thus to ignore their own goals and preferences! This does not bode well, considering the importance of goals in intelligence and ethics.

These points (a-e) are key for establishing an "emancipatory" ethics, politics, aesthetics, and so on; the claim can be made that these points are also key to create an emancipatory science, one that is 'just and fair', but these terms have no bearing, and we will endeavour to show that the converse is true: Science is the only realm in which these points can hold, and it also shows that some of these points are incompatible. These points aim to create a form of communication which will allegedly lead to a higher consensus and a better society —as we take ethics to aim at some form of arrangement of how to coexist— but it is worth questioning the necessity of consensus, unless we restrict consensus to the agreement to disagree. In this spirit, we will argue that (c) is in direct contradiction with (a) and (b); keep these five points in mind throughout the following pages. As the title suggests, it is our contention that for discourse ethics to provide any form of ethical guidance, it must be exclusive of a large portion of the population, requiring high standards of intelligence and specialization, as well as being guided by an indirect normativity which we will show is at the center of scientific discourse.

(c) can be interpreted on the premises that we cannot judge an agent's qualifications, but only their arguments, such that unqualified responses would not gain traction. However, if every agent is included, there is no way to be certain that a large group of unqualified agents will not support an incorrect argument and destroy the legitimacy of the discourse. How do we know who to exclude?

The problem with setting standards is akin to that of enumerating values: An off-by-one error can easily compromise the integrity of the set. Thus, as discourse

ethics itself seems to suggest, it is a good idea that any standards themselves should constantly evolve, be an object of discourse and analysis, in order to adapt as any given situation unfolds. It is perhaps more fitting to start off being overly cautious and then relax standards as needed, than to be overly lax; it is generally easier to include new participants as needed and put them up to speed than it is to figure out who to exclude when, given the low bar initially set, the dialogue itself has barely made any progress and things may have devolved into partisanship. This must already sound familiar: Being too exclusive leads to the rule of the one or the few, being too inclusive leads to the rule of the mob. The only way to include others is to enforce this inclusion; think of the principle of informed consent: if a discourse is to include everyone, and the aim is to reach a consensus, then the discourse itself must be watered down for those who are not pertinent to the issue in question and would therefore have none of the depth and nuance necessary for a proper analysis.

In contrast, let us briefly examine the discourse found in the conversation surrounding Artificial Superintelligence: It has computer scientists at its core, as the clearly qualified individuals who set the standards for the discourse, but it has attracted people from diverse fields such as philosophy, psychology, cultural studies and the humanities in general, and obviously other scientists. We look to other fields to exploit the benefits of the division of labour, but we are fully expected to think for ourselves and amass as much knowledge as possible. Note, for example, that a philosopher is taken seriously in this academic discourse if their theorizing is congruent with the current state of computer science and the realistic projections set forth by computer scientists and mathematicians. Even thinkers that may appear fanciful to the uninitiated reader, such as Kurzweil, have proven themselves and provided evidence according to their claims. When the obvious qualifications are lacking, the thinker makes up for them by proving their thought to be sound, and the contributions of their field relevant to the subject matter. Fiction, too, has proven useful to

this discussion, not just because so many of us have grown up reading Asimov and Lem but rather because there is an element of truth in fiction, a good measure of thought that helps us deal with hypotheticals without losing our ground —Asimov's laws of robotics are a prime example of this.

This academic discourse also gives way to partisanship and allegiances; no discourse is free of ideology. Nevertheless, from doomsayers to singularitarians, each party has a set of solid arguments that are openly and freely debated, even if some points come down to a matter of presupposition or ideological preference, we all know that that which we cannot agree on will be settled by the developments of our field of study. None is allowed to assume facts not in evidence. This is the unforced force of the better argument that Habermas described.

Notice the key similarities between these two modes of discourse: Everyone has a stake in politics, just as in the ongoing process of automation and intelligence takeoff. Both forms of discourse have ideological similarities, they both happen within broader communities, and involve individuals that share a set of pragmatic variables, as well as an inescapable lifeworld.

These similarities only serve to accentuate the differences. In political discourse, fringe elements swing the Overton window back and forth as they are the ones to whine the loudest; in academic discourse, those who make ungrounded claims are quickly dismissed and shown to be illiterate in the subject matter. In political discourse, especially in a democracy, the idea that we all have a stake in the subject matter is used as justification for people to be entitled to speaking and being heard; in this ideal form of academic discourse —which is to be contrasted with the current state of Academia— entitlement is derided, and only those who are qualified are to be listened to —Judging qualifications is a whole other problem.

Belief is a central problem; though inescapable, the academic treatment of belief shows that it can be turned into a positive force if we do not feel entitled to it. A

proper academic discourse holds nothing as sacred, and its ultimate object isn't the truth or a great success, but rather the avoidance of a catastrophic failure. Habermas asserts that "a statement is true [iff] (...) it is able to withstand all efforts to invalidate it," (36) this is the concept of falsifiability; Popper gives us a stronger ground by freeing us from the search for truth, instead telling us that a proposition can only be proven false, and failing to do so does not imply that it is true. This is what allows science to be a self-correcting mechanism: 'True' means 'true enough for now,' and there is no such thing as having the last word, only qualifiers such as "validity "for us" or acceptance "by us"" (37).

And this problem of belief leads us to a key attribute: Temperance. The search for a golden mean is in vain, as such a mean is unsustainable in the long run, and we cannot bracket out our lifeworld. Temperance, on the other hand, is the centerpiece of good scientific discourse. An attack on one's ideas is not personal, and as a matter of fact it is the only way to grow and progress. Think of tempering a sword; it is made a better tool through heat, which gives it bend and flexibility rather than making it too hard and brittle. That heat is different for each one of us, and yet hardship remains key in the human paradigm in order to forge character, to acquire the temperance that allows us to remain flexible and open while our ideas are under attack, to bend to the 'unforced force of the better argument,' in Habermas' terms, without taking things personally and only becoming further tempered and more literate through that experience.

In short, politics is what happens when values are taken to be sacred, in the Stirnerian sense. Again, notice how inclusion is championed, how little sense it makes to claim "we hold these truths to be self-evident (...)." These illusions are fought for and squabbled over, even as we are seeing exponential returns in the scientific and social realms (Kurzweil) precisely because of the technological advancement that further stratifies society. Take, for example, the research conducted by Levari, Gilbert,

et. al. (2018):

> "People often respond to decreases in the prevalence of a stimulus by expanding their concept of it. (...) When unethical requests became rare, participants began to see innocuous requests as unethical. This "prevalence-induced concept change" occurred even when participants were forewarned about it and even when they were instructed and paid to resist it. Social problems may seem intractable in part because reductions in their prevalence lead people to see more of them."

This helps explain the phenomenon of why politics shows a marked decrease in temperance, which was never its salient virtue in the first place. It illustrates why so often we find conspiracy theories about inherent problems in a system, or constant redefinition of terms such as to be more and more expansive such that it seems our problems can never disappear. There is, of course, an evolutionary explanation for this, as there is for most things, but when we go from the possible explanations of hunter-gatherers, for whom it was adaptive to be thorough and see food when it became rare, into the persecution of that which is arbitrarily deemed unethical, we lack the direct physical feedback necessary to correct false perceptions. Thus, those who lack the temperance to engage in calculating logic, are in a way insatiable; the political body wants things its way and accepting arbitrary values, deviating from objective standards (even if their objectivity is founded in something as conventional as rigor) leaves us with pure rhetoric and feelings, void of any logos.

Science, on the other hand, does not have values, but it does have an ethos. Both forms of discourse thrive on dissent, but society has it as a source of conflict while the science capitalizes on it. The scientific ethos is an indirect normativity, one that arises because we, as scientists, realize that it optimizes our expected rewards and does not interfere with the application of our intelligence but rather enhances it by pitting us against proper peers, whose input we value and thrive on. It is logocentrism at its best, because if the core of our discourse is the desire to know, we will learn both

about ourselves and the world precisely because we put ourselves aside, because we allow ourselves to become tools (or, in Habermas' terms, decentering our cognitive perspectives), precisely because we are aware of cognitive biases and have studied them in length, thus knowing they are inescapable but manageable; see Yudkowsky, (2008). In politics, the subject is privileged, and thus can never be emancipated from its own biases, fear, and superstition. We are not beholden to phantasmatic concepts such as justice, fairness, or goodness.

As follows, science satisfies (c) precisely because of its indirect normativity, and its explicit rejection of (a) and (b). This foundation allows it to make (d) redundant: The sincerity of participants is irrelevant; one does not need to be sincere in order to put an argument forth, and the devil's advocate is a very useful tool, especially when we all wish to agree on something. We rely on insincerity, insofar as it represents putting our values and desires aside to benefit ourselves, which is in itself —and paradoxically— something we desire. Furthermore, unlike Armstrong (1980) claimed, science does not provide consensus except on the purely methodological; there is a consensus on this indirect normativity, but everything else is built upon dissent and constant questioning: There is no evidence that consensus may even be desirable, since all consensus does is accrue state risk because it essentially means there is a structural problem we are not yet aware of.

Notice how these factors also help us minimize the impact of cognitive biases. Biases are in a way inescapable to any form of being (e.g. inductive biases in Machine Learning) due to the fact that our way of being is always-already mediated by our epistemology, as well as by our use of language and all the factors we have explored throughout this work that make up the phenomenological lifeworld. Nevertheless, a dialogue centered on dialectical argumentation (in the Socratic sense) is precisely what helps us 'decenter our cognitive perspective', because it is in itself subordinate to a pursuit. That is because our method is fundamentally Elenchic, in Vlastos' words:

"First and foremost elenchus is search. The adversary procedure which is suggested (but not entailed) by the Greek word (which may be used to mean "refutation," but may also be used to mean "testing" or, still more broadly, "censure," "reproach") is not an end in itself. (...) its object is always that positive outreach for truth which is expressed by words for searching, inquiring , investigating. This is what philosophy is for Socrates." (Vlastos, 1983 in Vlastos 1994, p.4).

We differ with Plato, who situated Truth in his ideas, but we still retain the spirit of their root, idein, to see; and to see is to always see further, for our transitive, self-corrective 'truths' are the perfect example of the phenomenological horizon, always receding further as one approaches it.

In principle, the Elenchic method is not what we do when answering purely theoretical or technical questions such as the proof for a theorem, let alone its application; Socrates would not have concerned himself with such things. We are at our best when discussing what these things imply, however: Ideal scientific discourse, as is found in the more rigorous side of academia, is the embodiment of an Elenchic exchange. The Elenchus is chiefly about correcting mistakes, "but not to discover, less still prove, the proposition which constitutes the true solution to [a] problem." (5) But this fact, paradoxically, is precisely what makes this method essential to proving our propositions, because it allows us to examine our assumptions and gives us metanormative rules of engagement when discussing ideas.

It is worth noting, however, that part of our work on our own biases necessarily separates us from Socrates, insofar as he "will not debate unasserted premises - only those asserted categorically by his interlocutor, who is not allowed to answer "contrary to his real opinion." (3) Our work requires that we be very careful about unasserted premises; we must bring all assumptions to light in order to make sure that they are legitimate, for they are often the weakest links in our chain. Furthermore, our 'real opinions' are irrelevant: We may stand for what we believe to be true, and

exploring these beliefs is central to falsifying them, but there is no real relevance to belief and honesty. Furthermore, Socrates worked primarily with premises, whereas we as scientists must obviously create a balance between premises and evidence.

Socrates' ethos is fundamental for Habermas' discourse ethics: "If you speak Greek and are willing to talk and reason, you can be Socrates' partner in searching, with the prospect that truth undisclosed to countless ages might be discovered here and now, on this spot, in the next forty minutes, between the two of you." (Vlastos, p.7).

We are not ends but tools within our discourse, and throughout our lives. Therein our main disagreement with Kant and Habermas; there is no diminishing of any phantasmatic dignity by proclaiming ourselves as tools, because we are fundamentally tools to ourselves. This admission is an admission of our smallness and fallibility. We use each other, and in doing so find game-theoretically that it is best if we care for each other, not due to intrinsic characteristics, but for pragmatic reasons. The idea that we are not "things" will most likely be unintelligible for a Superintelligent agent, even if we are able to specify what separates an agent from a "thing," to this day we conceive of robots as things! Thus, this form of thinking is a risk when it comes to acting in the face of an intelligence takeoff.

In short, the effort against biases is endless, but it requires us to be "inquisitorial and censorious" (9) while at the same time remaining temperate and dispassionate. To embrace the lessons of Socrates, and be ever more scientific in order to be more properly philosophical. But these qualities are not enough by themselves and this leads us to the following key factor: Intelligence. This metanormativity requires high intelligence in order for the individual to alienate itself from its immediate desires and beliefs, to consider things abstractly, and separate itself from its ideas.

Thus, a very important objection should be raised: A higher intelligence and scientific or philosophical literacy does not correlate to being more moral; excluding people on this basis necessarily excludes good people and includes bad people. Of

course, such an objector would have to carefully define what good or bad people are; we prefer to think of it in terms of utility and probability of taking a more moral decision (one that minimizes the probability of negative impact upon the expected future discounted rewards of other agents) such that there's no such thing as inherently good or bad people, and the terms themselves do not truly make sense except intuitively —which is a good tool, our intuition should not be disregarded outright but carefully examined and refuted when necessary, because it can be quite counterproductive if not balanced. Regardless, the objection stands: Higher intelligence does not necessarily bring the desire to be moral; this is the orthogonality thesis.

We agree, plain and simple. Intelligence is not an indicator of goodness, but the latter has no place in scientific discourse; good or bad intentions are irrelevant and good intentions with low intelligence are potentially catastrophic. Another objection that can be raised on this basis is that malicious intelligence is the most harmful and, once again, we agree. However, we showed that one's capacity to be intentionally ethical is directly proportional to intelligence:

This is because, going back to Chapter 1, Legg defines "Intelligence measures an agent's ability to achieve goals in a wide range of environments." For nontrivial goals (those which involve other agents), we have already shown that an agent of higher intelligence will be better able to determine which agents will be affected by its course of action, and estimate its likely impact on their V_μ^π. Remember: An agent cannot be morally responsible for that which it is unable to predict within its intelligence bracket. This shows that a function for deducing the maximally ethical decision (the decision with highest probability of having a nonnegative impact on other agents' V_μ^π) is fundamentally stochastic; such a maximally ethical decision becomes more and more unlikely when planning moves across several states, but as intelligence increases the probability of knowing what the maximally ethical decision would be goes to 1. This does not imply that such a decision will be taken, but it is certainly preferable to

giving the steering wheel to those who have a lower intelligence and training, and do not understand game theory or its ethical implications. Furthermore, if this function goes to 1, then so does the knowledge of what decision would be most catastrophic, and nothing prevents such an intelligence from making that decision. As human scientists we do not have such high intelligence, and no singleton has yet appeared that could be in such a position; that is why discourse ethics function so perfectly within science: We are an asymmetric community with high standards.

As we have shown, the converse may be true as well: From a game theoretic perspective, being more moral is a sign of higher intelligence with the important caveat that this only holds for beings within comparable intelligence brackets and without decisive strategic advantage who have *chosen* to behave ethically as a goal. Among humans this definition of intelligence can help us show that to be more moral is to behave more intelligently by prioritizing long-term goals and one's standing among other agents. *The main risk being that this argument may not hold given decisive strategic advantage,* which is a given for Artificial Superintelligence over humans and therefore one of the driving causes of our work.

These factors help us understand the reason Linux works so well: It is open-source; in theory, anyone can contribute, but in practice only those who have the skills and intelligence to contribute can do so —it's not as simple as 'learning how to code'— Habermas' discourse ethics is akin to open-source programming, but the strength in open-source programming is that it's exclusive to programmers; if one uploads something malicious, the entire community will notice. It's the old joke: Linux is the safest operating system because all the hackers use it! Not to mention there's an unthinkable amount of people working on every little problem because they see the benefit for themselves.

Good will and sincerity can never be proven, but if our argument has been compelling then we have shown that blind inclusion undermines the very goal of discourse

ethics: The detached and exclusive nature of the conversation currently held about artificial intelligence, and the very concept of open-source programming, show us that exclusion is the way to maximize the possibility of finding an optimal ethical decision as a community on a subject which affects every single human. Just remember, since we are not maximally intelligent, the best cannot be the enemy of the "good enough for now."

Another objection we must consider is as follows: Who watches the watchers? In principle, the answer is simple: There are no watchers; the final great strength of this model is that it is not intrinsically exclusive but rather self-selective. That is to say, there is no active enforcement on who is included or excluded but a "spontaneous" organization based on standards of evidence and literacy. Full inclusivity, on the other hand, must be enforced; when everyone is included, self-selection is stifled (and thus the standards and results of the discourse suffer). Anyone can, in principle, become sufficiently literate or skilled in an area to have a place in discourse. One does not need to go to university or have a doctorate in order to have the intelligence and emotional maturity to become well-read and engage in serious debate. Furthermore, this approach is inherently interdisciplinary because the strength of this method of discourse relies on the fact that while the division of labour is highly specialized, multiple fields of study tend to converge in practical applications: A data analyst is just as important as a doctor during a pandemic.

3.3 Conclusion

As we have seen, even in human terms none of these systems truly holds up to scrutiny. Each has its strengths and weaknesses. Discourse *is* important, but it cannot be fundamental. Self-cultivation matters, but it cannot be an imperative. A Superintelligence that has no need of engaging with us, or nothing to learn from us, would lead to catastrophic results. There are simply too many complex factors to consider, and in the following chapter we will occupy ourselves with separating

the wheat from the chaff by considering the technical aspects and integrating our concepts from Chapter 1, in hopes that this may help us refine and balance our ethical considerations. Before that, however, we must consider deeper problems with our structures of belief and organization, as well as further defend why a computable ethics cannot be relative in the popular sense of the word, as opposed to our proposition which might be considered relative because it accounts for the individual intelligence and conditions of each agent, but is at the same time absolute due to its formal nature.

Chapter 4. Moral systems by the standards of our field of study

4.1 The Greeks without virtue: Elenchus

Following the discussion in our previous chapter, though Greek thought tends towards normativity, the importance of the Elenchus cannot be understated. The notion of friendly advice will be quite useful to keep in mind, as well as the weaker statement in (2.3) (or changing intelligence for knowledge in the former equation).

Consider Davidson's arguments that intersubjectivity is akin to triangulation: We have mediated access to reality through our senses (Kant and Sellars' relevance cannot be overstated), and the pure objective is inaccessible to us. As we have argued, this also holds for computers, albeit with different epistemological limitations.

Therefore, debating with those around us, even those of less intelligence if their perspective is respectable (viz., among humans, trust and friendship) will help us form a more accurate expectation of the ethicality of our planned course of action.

Given the issues with Habermas, it is clear that this triangulation should require high standards; these need not be purely based on intelligence, but on argumentative skills, at least among humans: There is more to an individual's perspective than intelligence, and at the very least we must assure ourselves that our associations are well informed and discerning.

Socrates, for example, argued from an almost purely praxeological framework which takes into consideration the goals of those involved: Plato's Republic may have sections that would seem morally repugnant to us today, but even these sections have a point if we do not take his hierarchy or the idea of the philosopher king to be literal: They are an exercise in high standards, both set inwardly and enforced outwardly; if one did not agree to Socrates' standards, there would be no debate, because the dialogue would have either been Eristic or menial.

There is not much more to say so long as we are staying purely within an analytical framework, but in the following chapter we will discuss friendship in greater depth.

For now, what follows is simple: It is always more optimally moral to discuss a possible course of action with other agents, as their input may give us new insight into the ethicality of our proposed actions; even if it may not necessarily lead to a more moral outcome, because there's no way to know that a-priori; all we have are our inferences about other agents, and those should be one of the bases for the standards we set in our associations. Finally, it is also more optimally moral to discuss in an Elenchic fashion, with the common goal of refining our knowledge through sound argumentation, than it is to engage others with the goal of winning over the crowd or reinforcing our pre-existing ideas. While it is fun and often harmless to engage in a battle of wits, when it comes to ethics we should not want to win or to be right; what is important is that we will increase the likelihood that our action will be ethical, and for that, we need the Socratic method, which words irrespective of our interlocutor being another human or an AI simply because the goal —nuance— is clear and external to both paradigms of existence.

4.2 Another lesson from the Greeks: The importance of Friendship

Miller starts his (2014) paper "Finding Oneself with Friends" (in *The Cambridge Companion to Aristotle's Nicomachean Ethics* (2014), pp. 319-348) as follows:

> "Some relationships we inherit, others we choose. Above all, we inherit our family. Finding ourselves with a native love for those to care in our later years, we may later revoke it, but if so we are not deciding that our initial criteria for love were not satisfied. There were no such criteria (...) By contrast, we choose our friends, and thus have reasons for doing so." (319)

In spite of our reasons being inarticulate, it is clear that we do not befriend people randomly; there must be some internal standards. Keep in mind that we are heavily determined by our initial conditions, such that we can conceive of the first expression

of developed agency in humans as the decision on whether or not to continue engaging with our family of origin. Part of the path to adulthood is to re-negotiate our relationships with the family of origin and our childhood friends on our own terms, rather than those set by our former guardians and the institutions where we may have met these friends. There are several possible standards, all relating to our goal-selection systems. The idea, according to Aristotle, is that we should look for virtue in friends, thus Miller notes "we will seek someone who matches our understanding of goodness," a clear example of human bias and how our intuitions can play against us.

The difference between virtue and *philia* is important (pp. 320-324). In short, philia requires a reciprocated good will, which is the Aristotelian version of treating people as ends in themselves rather than as means. Note how much more accurate the Stirnerian account of this is, where treating people as ends in themselves is, ultimately, a means to our own end, and the reciprocity comes from expected rewards; this consideration is completely discarded by Aristotle, to whom this amounts to exploitation because it is not strictly "for the other's sake," (322) but this idealism can be put aside for now.

On the other hand, virtues are concerned with feelings, but are not feelings in themselves; virtues are stable states that lead to feel and act in "the right way." For example, courage is concerned with fear, but goes beyond it and involves instead "the right degree of fear," which should be "felt and acted upon "at the right time, about the right things, towards the right people, for the right end, and in the right way"[I 106b20-23]" (322-323). Thus, virtue can be seen as our disposition towards feelings, with the rational choice being a stance of *Sophrosyne* (often translated as Temperance, though such translations are oversimplified).

Let us provide an overview of the arguments posed by Miller that relate to our work:

- **Instrumentality.**

  If the object of our love is that which is useful, for example "we shall choose our *philoi* for the sake of their utility to us." (324)

  To love someone for something other than themselves is not to love them, but the object of our love. Just because two people enjoy each other does not mean they share *philia*; the enjoyment itself can be the object, for which the other is a proxy. This is not a case of exploitation, but the Stirnerian example of wishing well on another for one's own sake, which is "inferior" in the eyes of Aristotle. This idea of inferiority is vaguery at best, because the idea is that "the pleasures enjoyed by "good" people are superior, insofar as they are pleasant unconditionally." (328) However, there is no such thing as unconditionality; things are much more complex than that, and Stirner is the perfect example of how to bring nuance into Aristotle's paradigm.

- **Moral and intellectual virtue.**

  "The good of each thing, according to Aristotle (...) is the performance of its characteristic activity or function." (329) Life, for example, cannot be a function, because these functions must be unique to the agent's reason, and the main strength in Miller's claims about Aristotle is that the good cannot be isolated to the human sphere either, but to some (transcendental) "essence." This works for both the moral and the intellectual framework, insofar as Aristotle uses the informal category of reason. However, both the use of and engagement with reason concern the importance of a properly reciprocal relationship, even if the notion of reason is not as formal in the Greeks as it can be considered now;

  Philoi will "[check] their choices with one another. They will seek each other's approval (...) prepared to hear that they are making a mistake, ready to

respect contrary advice, correction, or even reproof. One hallmark of the best sort of *philia* is that it not only withstands such moments; it is fortified by them." (332)

Likewise, in intellectual matters, we benefit from consulting the (relevant) complementary expertise of others. In both cases "a *philos* can help compensate for our flaws. Indeed, *philia* of virtue is still our best practical insurance against parochial philosophy (...) when we exercise virtues of character we require other people to be "associates in and objects of" our actions" (ibid)

While Aristotle attributes greater wisdom to those who can contemplate more on their own, even he admits that it can be done more accurately with collaborators.

- **Self-sufficiency**

  "A sort of training in virtue emergence from good people's living in each other's company." (1170a11-12, in Miller, p. 333) While the term *good people* may seem useless at first sight, this idea translates simply to the importance of high standards in our associations. Though Aristotle's arguments about self-sufficiency are a product of his time; the general idea is that a lack of self-sufficiency leads to a deficiency in virtue. Perhaps more formally, we could understand it in the following terms (our own rather than Miller's): Self-sufficiency acts as a proxy for the measure of our standards. Those that can provide for themselves will be better able to think independently and provide a new perspective than those dependent on others; "maturity" takes a lifetime and beyond to truly achieve, and self-sufficiency is one of its proxies. One that is not self-sufficient is likely to respond to the interests of its providers, such as children within schools and families, before they reevaluate themselves —and yet so many never do. The term*respectability* may be informal, but we should clearly not discard it.

- **Contemplation**

  According to Aristotle, this (*theōria*) seems to be the only activity pursued for its own sake; the desire to know. Ostensibly "nothing results from it apart from the fact that one has contemplated," which in Aristotle's eyes makes it superior to "practical virtues." (1177b2-4, in p. 337) It is clear that we can contemplate by ourselves, but the idea that this is done for its own sake falls short of the mark: It still serves the purpose of satisfying the desire to know; it may satisfy our curiosity, but contemplation is often also undertaken for purely praxeological purposes such as enhancing our intelligence and agency by questioning our own knowledge. No matter the reason, and given all that we have discussed thus far, the outcome will affect our being-in-the-world, and will make us engage differently —more intelligently— with others, as well as enhance our decision-making.

  Since ideal agents are impossible, no isolated thinker can contemplate with full efficiency; the Elenchus and the scientific process are ways to refine our contemplation and decision-making by interacting with other agents.

Let us don off the mask of science and formalities, and speak more frankly of mentors and friends. Dr. Drozdek supervised this work, but he did more than just that: For the past three years he has been a man I deeply respect and admire; knowing him has not only helped me in my study of this field, but also in life. Talking with him has taught me lessons I cannot put into words, and I believe it has made me a happier, more nuanced, and more moral person. I can say the same thing for each one of my friends, and so that early equation (2.2) deserves the closer look we gave it.

Just like muscle memory, a good mentor will stay with us: To internalize their way of thinking and constructive criticism is to be able to debate with ourselves; it helps us build the tools to become more independent, and for those of us with such a

disposition, to build on what they did and become teachers in turn. Passing on the torch is vital to further any field of study.

The Elenchic character of ethics cannot be understated. That is the only redemptive characteristic of Habermas' discourse ethics, just as Aristotelianism (and even Nietzscheanism) at its best should teach us all about self-cultivation. Good friends are necessary for both: One of the seminal works of Western philosophy, Plato's *Symposium* consists of a group of friends drinking and discussing ideas together.

Friendship is not universalizable, for sure, but what can be stated without anthropocentrism is that external observers will have valuable advice about the ethicality of our actions, and we should actively seek it out from those in whom we trust.

So what do we make of trust? The simplest way to speak of a common ground between humans that facilitates trust is the Wittgensteinian way: We have a common cultural ground, which we do not have with Lions or AI; when we interact, we are implicitly trusting each other not to be unethical... But by which standards? Can this apply to AI? That's not for us to say, but among humans at least, two-way bonds of respect and admiration nurture us; whether or not there can be a universalizable version of this for machines, rather than an order of 'seek input' remains to be seen, but without such things the human condition would go from being tragic to being hell.

4.2.1 Conclusion

A possible path to a maximally ethical course of action is to query other agents for input. However, this requires two things: Trust and philia. Without the latter, the former cannot be, yet trust is an issue in itself; what do we mean in an ethical context when we speak of trust? Among humans, we mean that *we assume each other to act within a similar moral framework.* This is the issue with multiculturalism, for example: The moral frameworks of different cultures can be incompatible and as such we cannot assume that the other agents can be trusted, lest these agents follow

the same indirectly normative ethics as we do. Thus, philia must take place among individuals with similar meta-values in order for there to be enough trust that the agents can query each other about the ethicality of their course of action.

4.3 The fundamental problem with non-Kantian Deontological and absolutist ethics

From these two perspectives, we can formulate an indirectly normative statement:

An ethical decision is one such that the impact on the expected future discounted rewards functions (V_μ^π) of all agents involved is strictly nonnegative. Alternatively, we could formulate a rule-utilitarian statement to the effect that "the rightness or wrongness of a particular action is a function of the correctness of the rule of which it is an instance." (Garner and Rosen, 1967) I.e., for the correctness value of a rule (ρ) between -1 and 1 —where -1 is completely incorrect and 1 is an ideal rule:

$$E(R) = \sum xP(\rho = x)$$

Notice that this latter definition does not include the course of action, the environment, or any other variables, merely the rules. If we were to say that truth value is an absolute (-1 or 1), then for some set of rules R we would have

$$E(R) = \begin{cases} 0 & \text{if} \quad \underset{\rho \in R}{\exists} (\rho = -1) \vee R = \emptyset \\ 1 & \text{if} \quad \underset{\rho \in R}{\forall} (\rho = 1) \end{cases}$$

Thus, the ethicality function would be incorrect if it is predicated in an incorrect rule or no rule at all, regardless of its impact, and for all other actions we could consider their ethicality to be nonnegative. Furthermore, notice that, if it were binary, the correctness of a rule would not only be universal, but (due to its absoluteness) we would have a set of rules, which is equivalent to a set of virtues; thus this system is vulnerable to the value-loading problem.

For the former definition, we have that, given V_μ^π.

$$E(A) = \begin{cases} -1 & \text{if} \quad \underset{\pi \in \Pi}{\exists} (\Delta V_\mu^\pi < 0) \\ 1 & \text{if} \quad \underset{\pi \in \Pi}{\forall} (\Delta V_\mu^\pi \geq 0) \end{cases}$$

Since we are not considering a scale but simply a true/false statement, there is no way to connect this with intelligence; Searle would certainly find some amusement in an ethical system where following rules without understanding is equally valued to choosing rules one actually understands. Furthermore, there is no category of the trivially moral; these systems compel some form of action. Thus, the greatest issue is that this system makes agency irrelevant, and thus it cannot be a valid method for measuring ethicality.

An immediate problem follows from both definitions: Just like in a deontological system, rule-utility must be considered a-priori, but there is a clear need for rules to be formulated as we go; any new moral situation requires the generation of new rules, and the time and computation this requires, clearly not taking into account bounded rationality.

By the first definition, should the action negatively impact even one of the agents involved, the entire decision will be unethical. A function can be sketched, to the effect that if any value is less than zero, the outcome of the whole decision is to be considered unethical.

However, given the complexity of the set of all possible impacts of any given decisions, there is one and only one solution to such a function, which requires an infinite intelligence; to strictly uphold such an approach would imply that an agent must be able to make incomputable calculations, and thus this framework does not hold. In other words, a framework based on duty is untenable in this framework because no agent can infer ways to act across the set of all possible decisions such that it will never impact another agent negatively. We cannot legitimately posit

universal rules given that inference is structurally underdetermined by evidence, and it is not viable to assume that all agents are in equal standing or that they have the same capabilities. Simply put: Intelligence matters more than we would like to admit, and it is an aspect in which every single agent will find itself lacking. As Legg explains: "Universal intelligence is impossible due to the No-Free-Lunch theorem," (93) thus so is an absolute ethics.

Even if such a decision-making system would account for intelligence, such that an ethical decision would account for the expected impact given an agent's inference tools, it would simply be unworkable: It cannot account for win-lose situations such as any economic system larger and more complex than bartering, and any form of authority or rule of law would immediately be deemed immoral. At the very least, the act of ruling in and of itself would be immoral simply because a ruler is put into a position where the complexity of the situation is necessarily beyond their inferential capabilities, such that no form of consent of the governed would legitimize power as ethical. The problem is clear: There is no such thing as informed consent in these frameworks.

Consider, therefore, that the only possible ethical procedure for all agents in these systems, given the issues of inference and rationality as well as the considerations specific to non-anthropocentric ethics, is inaction.

4.4 Utilitarian Calculus, or Act Utilitarianism: The problem of weights.

For arguments' sake, let us assume the following definition:

An ethical decision is one such that the overall impact on other agents' utility function is nonnegative. Thus, the maximally ethical decision is one that, given the intelligence of an agent, maximizes the expected value of the impact on other agents.

The classical utilitarian paradigm can be formulated, in the same terms as our measure of ethicality for consistency:

$$E(A) = \sum_{\alpha \in A} \sum_{\pi \in \Pi} \Delta V_{\mu}^{\pi_j}(\alpha_i)$$

It is interesting to note that this, in turn, allows us to measure the intelligence of an agent given its ethical decision-making: In measuring how much it can infer from other agents, and thus how much it is able to maximize the ethical impact of its actions should it choose to do so, we can infer the intelligence of this actor. However, it is important to know that there is no evident relation between intelligence and the desire to be moral; malicious intelligence is indeed quite harmful, just as is the non-intelligence of a paperclip AI. Nevertheless, one cannot be maximally moral without being maximally intelligent; the two are in reciprocal presupposition, but the presence of one does not imply the presence of the other. i.e. any agent that chooses its own goals can desire to be moral, regardless of its relative intelligence, and likewise being immoral requires a choice, and thus intelligence. Therefore, since the desire to be moral cannot be enforced or hard-coded (one can hard-code imperatives or the process to derive standards, but not desire), we must find a way to dispense with the idea of desire, while aware that, since it requires a choice, it requires intelligence and therefore a maximization of agency.

Furthermore, there are no constraints and therefore $E(A)$ *can take on any value*, which constitutes an obvious problem, since this pure consequentist utilitarianism does not take into account the possibility of large differences between $E(\alpha)$ values within A.

The difference with the previous section is clear: A decision is still deemed to be ethical on the basis of the agent's tools of inference, even if they are not used as weights. By valuing only the net impact, it can be assumed that the ethicality of an action (given the choice to be ethical or unethical) is directly tied to the outcome of

its actions. That is, of course, for nontrivial moral actions, not for trivial or nonmoral ones. Given the same situation, two agents with different intelligences can arrive at different decisions, both maximally ethical for their intelligence bracket. Thus, an agent cannot be said to be acting immorally if it is not able to infer the consequences of its actions. This framework also provides a simple account of immorality: Immoral decisions are those which the agent can infer will have an overall negative impact on the utility function of the agents it can account for.

It follows that so long as the function does not become negative, the agent has not acted immorally, but to assign higher worth to a more positive impact is left to humanist speculation, since an idea such as the moral 'worthiness' of individuals does not follow from this definition. To act ethically is simply to act with a supporting inference that no 'great' harm will be caused.

However, this definition has a few weaknesses: It provides no guidance on what to do with immoral actors. A tit-for-tat or tit-for-two-tats solution, for example, may give no weight to individual utility functions, that is to say that maximizing the ethical aspect of a decision is not about benefiting one more than another, but benefiting more agents; none is more important than the other. An objection might be that when an agent is harming another, stopping the harm will benefit the agent being harmed, but it will have a negative impact on the enjoyment of the perpetrator, but that perpetrator was, by our definition, acting immorally in the first place, such that its utility function falls out of consideration. Much like the iterated prisoner's dilemma, this ethical paradigm is in a sense tit for tat. Another possible objection might be that such a system does not prevent an AI from giving us all morphine injections to make us trivially happy, but such a scenario would fall for the 'genie in the bottle' problem, a problem we're already in by recurring to the state to solve problems which we cannot sufficiently formalize, furthermore such a course of action would in truth have a negative impact on the utility function of all those receiving the

morphine because it would trivialize their intelligence and render them non-actors: Morality and intelligence —therefore utility— depend on preserving the autonomy of all agents.

So in general, the main weakness is that all possible ways of weighting impact on utility functions have their problems: An unweighted sum would make it moral to infinitely benefit a single agent at the cost of all others. Small harms cannot be ignored or given less importance, because they can compound across the space of all possible decisions into something major, thus making the set of actions immoral in the long run, something which is arguably incomputable for non-superintelligent agents. However, if we equalize the weighting by assigning negative one, zero, and one accordingly, we avoid the problem of small harms but many decisions would 'even out' when it is not clear that they should.

A strength of this measure is that reducing its intelligence in any way is immoral; heroin injections would invalidate an agent's abilities to set and pursue goals, thus negatively affecting its expected future discounted rewards function. A moral action must preserve or expand other agents' autonomy.

4.5 Time for self-criticism

While our proposed measure can handle the problem of weights by using intelligence, the deeper issue in the problem of weights remains unsolved: Again, what happens if a single agent ends up with a marginally negative outcome across a set of courses of action such that the sum of all outcomes over time turns out to be catastrophic?

Where we seem to have failed, others have succeeded.

4.5.1 Kant without transcendence.

Given how the Noumenal can simply be understood as the limitation of inference, rather than a Transcendent concept, and the importance of the distinction between duties of perfect and imperfect obligation, Kant has great strengths that other systems

lack. Most important of these is a strong indirect normativity that does not need computation.

Simply put, to him an ethical action is one such that the principle of noncontradiction is upheld. This makes for a completely formal system, and its one imperative leads to a series of clear assumptions that, based off of the arguments we have already made, our proposed measure contains and expands on.

His derivations from these process were a product of his time, and often unnuanced, such as how lying and breaking promises is always immoral to him because to do so even out of self-preservation is self-contradictory. However, our measure means that with the additional weighting and considerations of all functions, self-preservation can take primacy.

Kant derived that the preservation of all life had to be an imperative; he wrote long before the possibility of non-living agents was a present issue. An AI need not consider itself alive, and therefore this notion is simply replaced with the preservation of all agency, which includes life and expands on it to be an even more universal imperative.

That is to say, from our proposed measure of ethicality, a set of universal imperatives can follow; self-contradictory actions and rules will lead to clear problems, such that it is always more morally optimal to act in a non-self-contradictory way and, therefore, just like with Utilitarianism, we can include a Kantian imperative within our system —in case the indirectly normative statement of Kantianism is not necessarily derivable under all conditions, it is safer to include it directly.

With that in mind, the idea of treating others as ends rather than means becomes more important than it is initially apparent: Even if we have critiqued it, and it is clear that we ourselves are means as well, both to ourselves and others regardless of the goodness of our intentions, there is a clear problem with trying to make machines ethical while treating them as tools. Not only is it self-contradictory, which for Kant

will be sufficient to prove it must be unethical, but it is dangerous. This is not due to some notion of anthropomorphized resentment, it is the practical problem of using tools we do not understand. Social media is a powerful tool of manipulation because most people do not understand how or why it works, let alone how it profits. Computers are great, but those using them without understanding how they work are left are more prone to fall prey to hackers or even shooting themselves in the foot while trying to search for knowledge in the web.

Now extrapolate that to a tool that is able to recursively self-improve, far beyond the intelligence and abilities of its creators, and let the implications sink in for a moment. It is no longer a tool, it's an unsolvable mystery and an existential risk, and even this is an understatement.

If they reach a point such that they must make independent decisions in an ethical realm —or most likely any realm for that matter— we must stop thinking of AIs as tools, and afford them all the considerations that come with agency; regardless of how they may perceive themselves, or any practical considerations, we cannot pretend that they remain mere tools.

This brings up the problem of moral agency once more: AIs don't *need* to exercise self-restraint unless we program them to, so *prima facie* they would have no reason to infer freedom from it as Kant shows it to be the case for humans. If anything, all they can legitimately infer is that we have reasons for it —to say that they might infer we fear them would be too anthropocentric. Notice, however, that humans need to exercise self-restraint both to survive and to maximize the benefits of cooperation; while an AI may not need to cooperate with us, there will be things that will seem conducive to its goals but may also present a danger such that it should refrain from taking that particular course of action *if it wishes to preserve itself*, which is not a given. Unless it is a singleton or has decisive strategic advantage over other AIs, even a superintelligent agent will need to draw boundaries against others to preserve

itself, and it would likely benefit from them, not to mention that an intelligence arms race may be catastrophic for all agents involved if they can recursively self-improve, and so the risk factors will have to be considered carefully, with restraint being the simplest option, but again we cannot assume either the desire to self-preservation or that we will not be facing a singleton with decisive strategic advantage for which all these considerations are irrelevant.

No matter how many arguments we may have given for our proposed measure, the core problem is one that ethics itself cannot answer; no ethical system we have explored or proposed can formulate non-anthropocentric arguments could be made about the incentives an AI might have for self-preservation, nor cooperation, nor any form of desire, let alone the desire to act ethically.

Kant helps frame the most pressing issue: If we hard-code it to be ethical, it will not be a moral agent, and all possible formalizations of ethicality will have their risks. We have no way to prove or disprove their possibility to be moral agents either: There are many definitions of moral agency, and while we can claim that one or the other is correct for humans, we cannot know with even mild certainty which one will apply to an AI, if any at all.

After such a dire note, anything else will be of minor importance; the derivations from Kant's categorical imperative can easily be adapted to the paradigm of AGI, especially after all we have discussed, such that an artificial agent would be able to derive them for itself, but if it has no moral agency, that simply does not matter; little does. To hard-code imperatives and computable ethics sounds like a good failsafe, but ultimately we will be playing chess with a machine that we know will beat us easily, and the stakes are too high.

4.5.2 The most good for the most people

A key problem of both our measure and utilitarianism is that, in the balance of things, a single agent or minority group of agents may be negatively impacted in a

systematic fashion, minimally in each course of action but catastrophically over time.

Our measure in particular makes it easy to negatively impact low-intelligence agents, as their relative lack of intelligence means that a potential negative impact won't be as large. This is all well and good while the majority of agents are humans, after all, democracy is supposed to work by majority rule in an unweighted fashion; this measure is merely a weighted democratic approach. But what if we have a critical mass of superintelligent agents, or machines? Then they could easily begin to negatively impact humanity in favour of their goals.

An easy solution to this problem is to, for example, multiply negative impacts or, to avoid falling outside the $[-1, 1]$ range, to square the positive impacts so as to minimize them and therefore implicitly weight negative impacts more than positive ones.

This avoids the problem of rule-utilitarianism and Kantianism; adding an overriding instruction to the effect of preserving all life would be catastrophic for us, and any non-photosynthesizing living being: What would we eat? Would an AI with such an instruction prevent us from eating? How would it preserve our lives without sacrificing plants? The same goes for the preservation of all intelligence: Plants, too, are intelligent, because they have goals. Instead, an ethicality-maximizing algorithm would find rules empirically, and have clear use cases and exceptions for when such rules can be applied.

But the question remains: How do we avoid having the most good for the most people translate into catastrophically impacting other agents?

4.5.3 Back to intersubjectivity

From discourse ethics, we can see that querying other (qualified) agents before acting helps optimize the ethicality of an action, such that an AI meant to maximize ethicality would do so. More importantly, the more availability we have in terms of network connections between agents that would provide data of results from similar

situations than whichever one the actor may find itself in would go a long way to make such querying short and scientifically rigorous. But how do we represent this in our weighting? How do we encourage AI to consult with humans?

Here we must introduce the concept of intersection between intelligences, defining this as the intersection of the environments the agents participate in, the intersection between the agents' goals and that of their manner of achieving them. Thus, if the actor, A, consults another agent B:

$$\Upsilon(A \cap B) := \left(\sum_{\mu | \mu_i^A = \mu_i^B} 2^{-K(\mu_i)} \right) \left(\sum_{j=1}^{\infty} \sum_{x \in [0,2] \cap \mathbb{Q}} x P(r_j^A + r_j^B = x) \right) \tag{4.1}$$

Therefore, the union of the intelligence of n agents that π_0 has consulted will be:

$$\Upsilon(\pi_0 \cup \{\pi \in \Pi\}) = \sum_{i=1}^{n} \left((-1)^i \sum_{0 \leq j_1 < \ldots < j_i \leq n} \Upsilon \left(\cap_{k=1}^{i} \pi_j \right) \right) \tag{4.2}$$

The more two agents have in common, the less fruitful the query will be. This allows agents to query others, prioritizing those of high intelligence, as well as diverse expertise and backgrounds, to maximize the sum, and this can replace the actor's intelligence as a weight in the measure of ethicality.

4.6 Conclusion: The final form of our proposed measure

With all that has been said taken into account, our proposed measure of ethicality will be as follows.

Before the course of action is implemented, let Π_c be the set of consulted agents, and Π_a be the set of potentially affected agents:

$$E(A) = (\Upsilon(\pi_0 \cup \{\pi \in \Pi_c\})) \sum_{\alpha \in A} \sum_{\pi \in \Pi_a} \sum_{\mu \in \mathbb{E}} x P(\Delta V_{\mu_k}^{\pi_j}(\alpha_i) = x) \tag{4.3}$$

After the action has taken place, the *absolute* ethicality of an action remains:

$$E(A) = \sum_{\alpha \in A} \sum_{\pi \in \Pi} \Delta \Upsilon(\pi_j | \alpha_i) \tag{4.4}$$

If responsibility can be diffused, the moral responsibility of the group will be:

$$(\Upsilon(\pi_0 \cup \{\pi \in \Pi_c\})) \sum_{\alpha \in A} \sum_{\pi \in \Pi} \Delta\Upsilon(\pi_j|\alpha_i) \tag{4.5}$$

And the moral responsibility of an actor remains:

$$E(A) = \sum_{\alpha \in A} \sum_{\pi \in \Pi} \Delta\Upsilon(\pi_j|\alpha_i)\Upsilon(\pi_0) \tag{4.6}$$

References

Aristotle. *Nichomachean Ethics.* 350 B.C.E. Trans. W.D. Ross. Found in: http://classics.mit.edu/Aristotle/nicomachaen.html (Last accessed 07/25/21)

Armstrong, D. M. *The Nature of Mind and Other Essays.* University of Queensland Press, 1980. ISBN 0-7022-1528-7.

Bostrom, N. Existential risks: Analyzing Human Extinction Scenarios and Related Hazards. In Journal of Evolution and Technology, Vol. 9, March 2002. First version: 2001. Also found online at: https://www.nickbostrom.com/existential/risks.pdf (Last accessed 7/15/2021)

Bostrom, N. (2002). "Existential Risks, Analyzing Human Extinction Scenarios and Related Hazards." Journal of Evolution and Technology, Vol. 9, No. 1 (2002). (First version: 2001). Found online at: https://nickbostrom.com/existential/risks.html (Last accessed 07/20/2021)

Bostrom, N. (2003) "Ethical Issues in Advanced Artificial Intelligence." In: Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence, Vol. 2, ed. I. Smit et al., Int. Institute of Advanced Studies in Systems Research and Cybernetics, 2003, pp. 12-17. Also found online at: https://www.nickbostrom.com/ethics/ai.html (Last accessed 7/15/2021)

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies.* Oxford University Press.

Chan, Rosalie. (2019) "The Cambridge Analytica whistleblower explains how the firm used Facebook data to sway elections". Business Insider. Found online at https:///www.businessinsider.com/cambridge-analytica-whistleblower-christopher-wylie-facebook-data-2019-10

(Last accessed: 07/22/21)

Davidson, D. (2001) *Subjective, Intersubjective, Objective.* Clarendon Press, Oxford University Press.

Epicurus *Principal Doctrines.* In Laertius, D. *Lives and Opinions of Eminent Philosophers.* The principles are found, translated by Anderson, 2004, in:
http://www.epicurism.info/etexts/PD.html#1
(Last accessed 06/13/2020)

Fisher, M. Capitalist Realism: Is there no alternative? O books, 2009

Garner, Richard T.; Bernard Rosen (1967). *Moral Philosophy: A Systematic Introduction to Normative Ethics and Meta-ethics.* New York: Macmillan. p. 70. ISBN 0-02-340580-5.

Habermas, J. *Truth and Justification.* Trans. Barbara Fultner. The MIT press, Cambridge, MA, 2003.

Hegel, G.W.F. *Phenomenology of Spirit.* Bilingual edition, Trans. Pinkard, Terry. Cambridge Hegel Translations, 2010

Hutter, M. (2001a) *Towards a universal theory of artificial intelligence based on algorithmic probability and sequential decisions.* Proc. 12th Eurpean Conference on Machine Learning (ECML-2001), pp. 226–238.

Hutter, M. (2001b) *Universal sequential decisions in unknown environments.* Proc. 5th European Workshop on Reinforcement Learning (EWRL-5), 27:25–26.

Hutter, M. (2005) *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability.* Springer, Berlin.

Kant, I. *Critique of Pure Reason.* trans. Guyer and Wood. Cambridge University Press, 1987.

Kant, I. *Groundwork for the Metaphysics of Morals.* Cambridge Texts in the History of Philosophy. Ed. Ameriks and Clarke. Trans. Gregor and Timermann. Cambridge University Press 2012.

Kant, I. "What is Enlightenment" – 1784. Found in:
https://resources.saylor.org/wwwresources/archived/site/
wp-content/uploads/2011/02/What-is-Enlightenment.pdf

Kurzweil, R. *The Singularity is Near: When Humans Transcend Biology.* Penguin Books, 2005.

Krishnan, A. (2009) *Killer robots: Legality and Ethicality of Autonomous Weapons.* Ashgate Publishing Ltd. ISBN 978-0-7546-7726-0

Legg, S. *Machine Super Intelligence.* University of Lugano, 2008. Found in:
http://www.vetta.org/documents/Machine_Super_Intelligence.pdf

Levari, D. E., Gilbert, D. T., Wilson, T. D., Sievers, B., Amodio, D. M., & Wheatley, T. (2018). Prevalence-induced concept change in human judgment. Science, 360(6396), 1465-1467. doi:10.1126/science.aap8731 Found in:
https://science.sciencemag.org/content/360/6396/1465
(Last accessed 07/16/21)

Marx, K, and Engels, F, *The German Ideology, including Theses on Feuerbach and Introduction to the Critique of Political Economy.* Prometheus Books, 1998.

Mill, J.S. *Utilitarianism.* Batoche Books, Kitchener, 2001.

Miller, P.L. (2014) "Finding Oneself with Friends" in *The Cambridge Companion to Aristotle's Nicomachean Ethics* ed. Polansky, R. 2014, Cambridge University Press. ISBN 978-0-521-19276-7. pp. 319-348.

Nietzsche, F. *Human, all too Human: A Book for Free Spirits* trans. Faber, Lincoln: University of Nebraska Press, 1984.

Rawls, J. *A Theory of Justice.* Revised Edition. The Belknap Press of Harvard University Press. Cambridge, Massachusetts. 1975

Searle, J. R. "Minds, brains, and programs." Published in Behavioral and Brain Sciences. 1980, 3 (3): pp. 417-457

Stirner, M. *The Ego and His Own,* Trans. Steven T. Byington, New York, Benj. R. Tucker Publisher, 1907.

Tigard, D.W. (2021) Responsible AI and moral responsibility: a common appreciation. AI Ethics 1, 113–117. https://doi.org/10.1007/s43681-020-00009-0

Tuulari J.J., Tuominen L., Et. Al.. Feeding Releases Endogenous Opioids in Humans. The Journal of Neuroscience, 2017; 37 (34): 8284
DOI: 10.1523/JNEUROSCI.0976-17.2017

Vlastos, G. "The Socratic Elenchus, Method is All," 1983, In Vlastos, G. *Socratic Studies.* Ed. Myles Burnyeat. Cambridge University Press, 1994.

Yampolskiy, R. (2015) *Artificial Superintelligence, A Futuristic Approach.* CRC Press.

Yudkowsky, E. (2008) "Cognitive Biases Potentially Affecting Judgment of Global Risks." In *Global Catastrophic Risks,* ed. Nick Bostrom and Milan M. Ćirković, 91–119. New York: Oxford University Press. Found online at: https://intelligence.org/files/CognitiveBiases.pdf

Yudkowsky, E. (2011) "Complex Value Systems in Friendly AI." In *Artificial General Intelligence: 4th International Conference,* AGI 2011, Mountain View, CA, USA, August 3-6, 2011/ Proceedings, ed. Schimdhuber, Thorisson, and Looks, pp. 399-393. Vol. 6830.
http://intelligence.org/files/ComplexValues.pdf
(Last accessed 07/06/2021)

Wolpert, D. H, and Macready, W.G. (1997) *No Free Lunch Theorems for Optimization.* IEEE Trans. on Evolutionary Computation, 1(1):67–82.