UCLA School of Law

Law & Philosophy Program

Legal Theory Workshop
UCLA School of Law

Ruth Chang

University of Oxford

"Aligning AI with Human Values: A Proposal"

Thursday, April 17, 2025 1:10 – 2:40pm
Law School Room 1314

Draft, April 2025. For UCLA Workshop.
Please don't cite or quote without permission.

Aligning AI with Human Values: A Proposal

Ruth Chang
ruthechang@gmail.com

Abstract:
Arguably the most important open problem in the development of Artificial Intelligence is the 'Alignment Problem', roughly, how do we ensure that AI processes and outputs align with our best considered judgments about what is and is not valuable? Current strategies for achieving alignment, ranging from regulation to development interventions such as reinforcement learning from human feedback, prompt engineering, data scrubbing and the like, have led to spectacular failures and fall well short of achieving machine-human value alignment. Proponents of such strategies hold out the hope that more data, more reinforcement learning, more monitoring, more prompt engineering -- and more environmentally unfriendly computing power -- will eventually allow current strategies to achieve alignment.

Common to these approaches is the strategy of 'fixing' AI after it has been created. Instead of imposing ad hoc *guardrails* ex post, *we suggest a more radical approach: reexamination of the fundamentals of AI design. We identify two fundamentally mistaken assumptions about human values made in current AI design, which suggest that i) no amount of data scrubbing, prompt engineering, reinforcement learning from human feedback and the like can achieve alignment, and ii) an alternative AI design that corrects for these two mistakes is a* sine qua non *for achieving value alignment. We propose a conceptual framework, the 'Parity Framework', which avoids these two mistakes about values. It has three distinctive features. It is a 'values-based' approach to AI design according to which values, rather than their non-evaluative proxies, are the inputs for machine processing. It recognizes that machine-human value alignment requires development of 'small ai', that is, algorithms that carve at 'the normative joints,' – different*

algorithms for sets of choice situations in which what matters in the choice is the same. And it puts the human in the loop in a distinctive place – in genuine hard choices. Human feedback in hard choices in turn reflects the evolution of human values through human response in such choices, provides explainability where it most matters, and helps to create a machine-human hybrid intelligence that can provide a foundation for machine-value human alignment.

Arguably the most important open problem in the development of Artificial Intelligence, understood broadly to include any machine technology that attempts to do what humans do when they do or attempt to do things, is the *Alignment Problem*: roughly, how do we ensure that AI outputs align with human values? Indeed, achieving value alignment between humans and machines might allow us to get 'for free' at least a partial solution to the other looming problem in AI development that has received far more attention, the *Control Problem*: roughly, how do we ensure that AI outputs are safe, especially from a wide, socio-technical perspective? If AI aligns with our values, then since human flourishing is a human value, perhaps aligned AI might avoid imposing undue risks on our well-being, becoming our overlords, or, worse yet, extinguishing us altogether.[1] Solve the Alignment Problem, many experts agree, and the prospects for genuine human flourishing in a future filled with AI begins to look reasonably rosy.

So far, there are two main paths to achieving machine-human value alignment. One is regulation.[2] On August 1, 2026, the EU AI Act, which regulates AI according to its supposed riskiness, will be officially enforceable against non-compliant technologies (with August 1, 2027, being the drop-dead date for 'high risk' AI). The AIA is widely recognized to be the most advanced regulation of AI in existence, but it remains to be seen whether other nations will adopt its comprehensive approach, or whether the EU has boxed itself into a regulatory regime that will leave it significantly behind in the technology race. In the United States, comprehensive AI regulation looks increasingly unlikely. Just about a month ago, Trump revoked Biden's rather innocuous AI Risk Management Framework and Blueprint for an AI Bill of Rights, which mostly promulgated very general principles for AI development with which it is difficult reasonably to disagree.[3] The United Kingdom appears to be taking a somewhat splintered, decentralized approach. On the one hand, it gives individual government regulators responsibility over regulating technologies in their domain of operations, with a central authority whose power to

coordinate rules and regulations still an open question. On the other hand, the UK appears to be trying to position itself as a global leader in AI safety, as the host of the Bletchley AI Safety Summit held November of 2023 and the recent creation of independent AI Safety Institutes.[4] The professed aim is to manage risk while ensuring that regulation does not stifle innovation.[5] China takes a decidedly centralized approach to AI regulation. It requires all technological products to be approved by the government. It, too, has focused on AI risk – to everything from the mental health of minors to socialist values (but not individual freedoms).[6] Other countries have regulations in the works, too.[7]

Can regulation do what's needed? Ascertaining in advance the potential risks of a new technology is notoriously difficult. Who would have thought in the early days of Facebook that it could be used as a platform for insurrection? Moreover, regulation's typically long lag time makes it a sluggish and blunt tool for responding to the dynamic and finely-drawn demands of alignment in a rapidly changing socio-technical reality. Indeed, the EU AIA will have taken five years from soup to nuts. Finally, there is the problem of knowing which regulations will have what effects on AI development, output and uptake. As Rishi Sunak put it, "How can we write laws that make sense for something we don't yet fully understand?"[8] Regulation, while very much needed, will likely consign us to a game of whack-a-mole, especially if, as we suggest here, the Alignment Problem arises because of deep problems within AI design itself.[9]

Another path taken on the road to alignment is offered by technologists. Computer scientists and engineers have produced a suite of interventions in AI development that aim to make machine outputs better align with human values. These include monitoring, data scrubbing, prompt engineering, and fine-tuning, especially reinforcement learning from human feedback (RLHF), where humans or human-trained algorithms evaluate multiple machine outputs to train the model to give more human-aligned outputs. But these strategies, while boasting of some successes, are plagued by spectacular failures.[10] Alignment, at present, is still very much out of reach.

We are at a fork in the road. Pessimists think that no amount of regulation or development intervention will allow us to achieve our goal of alignment. You can try to install legal guardrails on the development of AI, and you can throw as much data, compute, and RLHF as you like at algorithms, but they will still fail to align with human values. Optimists put their store in hyper-scaling. With much more data and computing power, and ever more subtle RLHF and fine-

tuning, we will eventually achieve alignment.[11] There is, at present, a theoretical stalemate. In the meantime, Big Tech are investing huge resources – at what looks to be a hefty environmental cost – in a hurry to prove the optimists right.

Might philosophers help? Might there be a relatively simple *philosophical fix* – some philosophical insights about values, for instance – that if applied to technological design would go some way – perhaps a long way – to solving *Alignment?*

Here's a hypothesis. Progress in addressing the alignment problem has been stymied because technology is currently designed on the basis of two fundamental mistakes about human values, mistakes that must be addressed for there to be any hope of achieving alignment and control. Regulation and development interventions will fail to provide a solution because they are *ad hoc* guardrails limited in their ability to prevent misaligned AI from being created in the first place. Instead of approaching *Alignment* by imposing constraints *after* AI has been designed, we need a more radical approach: a reexamination of the fundamentals of AI design. In particular, if we are to achieve alignment, we need a framework for AI design that avoids the two mistakes about human values made by current AI design. I propose such a framework and sketch an axiologically-grounded but computationally possible design model – The Parity Framework – for value alignment between humans and machines.

I.

What is the Alignment Problem?

It will help first to get clearer on the alignment problem. The locus classicus is given by Norbert Weiner's admonition some sixty years ago:

> "If we use, to achieve our purposes, a mechanical agency with whose operation we cannot interfere effectively…we had better be quite sure that the purpose put into the machine is the purpose which we really desire".[12]

Weiner called this the problem of 'cross purposes' – a problem that arises when the machine's purposes are different from the purposes for which we designed it – but modern-day scholars have dubbed it the problem of 'alignment'.[13]

Modern day technologists seem to follow Weiner in his statement of the problem. But Weiner's statement, and the prevailing understanding of *Alignment* among technologists, makes three problematic assumptions.

First, there's the assumption that the purpose for which we build a machine is the very same purpose we should put into the machine. Philosophers have taught us, however, that some purposes should be pursued not directly but indirectly. If your purpose is to be happy, you had better not aim at happiness but at having good friendships, health, education, etc., -- the things that constitute a happy life. Moreover, in thinking about what purpose to put into a machine, we need to be mindful of the fact that machines have different capacities and constraints which affect how they can understand and execute a purpose.[14] If my purpose is to change a chandelier lightbulb, I need to get a ladder. Robert Wadlow, the tallest man in the world standing at 8 feet 11 inches, would only need to raise his arm. Machines, like Wadlow, are differently bounded from ordinary humans, and so the way they approach a purpose may be different from the way we do. In short, instead of blithely assuming that the purpose for which we design the machine is the purpose we must put into the machine, we need to reflect carefully about which purposes we should put into the machine in order to achieve the purpose for which we design the machine.

Second, there is the assumption that AI alignment is a matter of matching AI outcomes with what we humans *want,* or as Weiner puts it, what we "*really* want." But desires or preferences are not the right metric for assessing alignment with human *values.* This is because what we prefer is one thing and human values are another. It may be a fact about Adam that he always prefers to look out for No. 1, maybe because he's an unapologetic narcissist. If we subject his preferences to purely formal constraints to determine what he 'really' wants, a la Bernard Williams, such as cleaning up his logic, correcting his false non-normative beliefs, ensuring coherence among his mental states, endowing him with full imaginative non-normative powers, and so on, his narcissistic preferences could remain. His preferences, though, would not reflect our human values. If, in the alternative, we subject his preferences to *evaluative* constraints and understand him to be what philosophers call 'substantively rational', which includes having the capacity to recognize and respond to human reasons and values, then his preferences could in principle reflect human values. But preferences cannot be guaranteed to reflect human values without being themselves subject to and constrained by human values. In this case, preferences are mere downstream effects of responsiveness to human values, and it is the values themselves

that give preferences the needed imprimatur. In short, we should not be attempting to align machines with our preferences unless they are reflective of human values. But there's no guarantee that preferences reflect human values unless that are constrained by those values. So there's no avoiding human values in a metric for alignment. This is not a surprising conclusion. Whether machine processes and outputs align with human values can only be determined by human values.[15], [16]

There is a third problem with Weiner's statement. It implicitly suggests that there is just one alignment problem, and this suggestion has led some technologists to suppose that by solving one problem in alignment, they have solved them all. I believe that there are (at least) two distinct problems falling under *Alignment* that should be distinguished. Indeed, it is by distinguishing these two alignment problems that we can uncover the two mistakes about human values embedded in current AI design.

So how should we understand the Alignment Problem? We might propose the following definition:

> *Alignment* is the set of problems involved in ensuring that machine processes and outputs align with our human best-considered judgments about what is good/bad/right/wrong/just/ egalitarian/fair/unbiased/tawdry/funny, and so on – [insert any human value you like here].[17]

This is not a reconceptualization of the problem but a more accurate statement of the problem that technologists understand themselves as trying to solve that also encorporates what philosophers already understand about human values.

One further distinction is worth making. There are two 'kinds' of AI, distinguished by the kind of purpose for which it is designed. Sometimes we design an AI to achieve purely non-normative aims. We might want a machine to determine which of two business plans will yield the highest profit, which bail decision will lower the probability of recidivism, which suite of medicines will save the largest number of lives, or which algorithm will most accurately determine tumors as malignant. Call this 'Non-normative AI'. It is what we might intuitively think of as purely calculative AI that has no truck with values or normative aims.

Other times we design AI to achieve, at least in part, evaluative or normative aims, such as finding the *best* candidate for the job, a *morally good* way to avoid war, a child foster care assignment that is in the *best interests* of the child, a *fair* prison sentence, and so on. Call this

'Normative AI'. The Alignment problem is only a problem for Normative AI, AI designed to achieve at least some evaluative aims. Of course, Non-normative AI can be used for nefarious purposes – you can use a calculate to figure out how to kill as many people as possible – but that is not a problem of aligning the *machine* processes and outputs with human values. Rather it is a problem of getting humans to use those outputs in ways that align with human values.

Note that these days most AI is Normative. This is because even if an AI is designed to achieve primarily non-normative ends – e.g., to accurately identify malignant tumors – we also design it to achieve those ends in a way that does not run afoul of certain values – e.g., by discriminating against minority patients. In general, we build our algorithms to be safe, effective, and fair. These are evaluative ends. So *Alignment* is a problem for the vast majority of modern AI systems.

II.
Two Alignment Problems

There are two distinction alignment problems. Examining each will uncover, respectively, a mistake about human values embedded in current AI design.

A. The Covering Problem

The first is the *Covering Problem,* the problem of determining the criteria that 'cover' the purpose and only the purpose we put into the machine. Often our purpose is one thing, but we end up designing the machine to do another. Following Weiner's invocation of the fable of King Midas, Stuart Russell calls this the 'King Midas problem': "You get what you asked for, not necessarily what you want."[18] We need to make sure that the criteria we give a machine in fact cover our purpose or aim and no other purposes or aims. For convenience we can treat this purpose as evaluative, even though it may be combined with certain nonevaluative purposes.

A notorious case of covering failure comes from Amazon, which developed a recruitment algorithm whose purpose was to deliver the best candidates for any position among thousands of applicants. The criterion used to cover the 'best' was 'match with successful employees hired in the past'. Since men dominated in the sheer number of applications and therefore in the number of those actually hired, the algorithm naturally trained itself to see a 'best' candidate as male,

downgrading resumes that contained qualifications such as 'women's chess club captain'.[19] What Amazon wanted was the 'best', but what they asked for was 'people most like those we've now got on the books.'

Or consider the algorithms used to manage heath care costs by identifying "high risk care management" individuals, that is, those whose health care needs contribute to the lion's share of total health care costs in the U.S. Some estimate that 5% of people account for over half of total health care costs in the U.S.[20] By identifying the sickest such individuals, interventions could be made to prevent or mitigate their disease and thereby minimize overall health care costs, which in the U.S. outstrips that of any other country by almost double per capita.[21] In a careful piece of sleuthing, Obermeyer and colleagues discovered that the algorithm, trained to predict the actual medical costs of individuals over their lifetimes, excluded more than half of the Black people who should have been identified as high risk. Black people were underrepresented as targets for intervention relative to white people because the training data reflected the fact that they tend not to receive the same medical care unless they are much sicker, and so a prediction of their lifetime healthcare cost obscured the severity of their disease.[22] While it seems natural to use the criterion 'health care costs over lifetime' to cover 'who needs the most expensive health care interventions over their lifetime', this criterion covers only what people are projected to spend on health care, not how sick they are when they receive that health care.

The Amazon hiring and the health care cost algorithms are paradigmatic examples of how tricky it is to find *covering criteria* that cover all and only the ends for which we design the machine.

To a philosopher, the Covering Problem may seem intractable. Aristotle pointed out that it is a mistake to think that one can list, *ex ante,* all the evaluative factors that may be relevant to achieving a given evaluative end in all possible circumstances. And some of our key normative concepts, like *justice,* are arguably 'essentially contested' – their application is always subject to reasonable contestation.[23] And so, Aristotle thought, we need *phronesis,* that is, practical judgment or wisdom to be exercised in particular circumstances and contexts. As we will suggest, Aristotle was importantly right about the need to consider evaluative ends in context-sensitive ways.

Computer scientists have developed some clever ways to address the Covering Problem. In machine learning, the most interesting strategy to my mind is Stuart Russell's 'cooperative

inverse reinforcement learning' (CIRL).[24] The innovation of this approach is to build uncertainty into the machine's reward function and to reduce this uncertainty over time by, for example, machine-human game playing or by the machine asking the human a battery of questions. The hope is that such machine-human interaction will allow the machine to learn the human's reward function and consequently the criteria that supposedly cover the purpose for which the algorithm was designed.[25] But there are two problems with the approach.[26] First, at some point the machine will deem itself as having completed its learning. It remains to be seen whether what it has in fact learned matches what we want the machine to do. Second, and relatedly, programming in CIRL has yet to figure out how to catch the rational parameters of a human response. If you tell the machine that you want to go to the store, there is always the possibility of misunderstanding – e.g. that you want to go there every day – unless you give it some restrictions. But it's unclear how in a two-person game we can give a machine all the complex, multitudinous restrictions required to correctly identify every goal we might have.[27]

There is a deeper problem. We now come to the First Mistake about human values embedded in current AI design. Along with all current AI design, CIRL makes a critical but mistaken assumption about human values. Call this 'Values Proxy':

> *Values Proxy:* One can always achieve an evaluative end, V, by pursuing instead a nonevaluative proxy, P, across a wide range of circumstances.

Indeed, current algorithms for machine learning are systematically designed to achieve evaluative ends through nonevaluative proxies.

But *Values Proxy* is mistaken. The fable of King Midas shows us why. What are the King's evaluative ends? He wants certain values and the goods that instantiate them: the love of his daughter, good health, happiness, knowledge and understanding, rewarding friendships, worthwhile achievements, and peace in his kingdom. His mistake, like that of current AI design, is to assume that these values and goods can be achieved by getting something non-evaluative instead: gold.[28]

Thus, the First Mistake AI design makes about human values is to assume, along with King Midas, that we can use nonevaluative proxies for our evaluative goals across a wide range of circumstances. Such proxies will necessarily fail to 'cover' the evaluative aims for which we design AI in the first place. Different circumstances will require different proxies and different

relations among those proxies. In this way, current AI design, which assumes *Values Proxy*, guarantees value misalignment in many cases.

We need to design AI that avoids using nonevaluative proxies to achieve evaluative goals. One way to do this is to adopt a 'values-based' framework for AI design. This requires a very radical shift in how we think about AI. If we are to achieve our evaluative ends, we should build AI to process *evaluative*, not *nonevaluative* data. Evaluative data would be information about *values,* that is evaluative facts most naturally accessed through considered evaluative *judgements.* We return to 'values-based' AI design at the end of the essay.

B. The Tradeoff Problem

Normative AI is typically designed to achieve multiple evaluative purposes or ends. The best job candidate will be loyal and sincere, supportive of his colleagues, able to think creatively about problems, productive, and so on. Even when a machine is built to achieve a single evaluative purpose, that purpose will typically have multiple aspects. So now we face a problem: how should we trade off multiple criteria or aspects? This is the second alignment problem, the *Tradeoff Problem*.

Standard technological design attempts to address the Tradeoff Problem in one of two main ways.[29] One is to fold the Tradeoff Problem into the Covering Problem and treat them as a single problem. As we saw in the case of Amazon, machine learning algorithms employ a single nonevaluative criterion – e.g., 'actually hired' – to run as a proxy not only for being the best hire but for tradeoffs between, say, productivity and team-spiritedness. We've already suggested that no non-normative criterion could work as an accurate proxy for value across a wide range of circumstances. Misalignment in value tradeoffs is built right into this form of AI design.

The other strategy, more commonly employed in symbolic systems, recognizes that there are multiple criteria that determine output, assigns weights to these criteria in the program, and then evaluates the output. If the output looks amiss, the weights are adjusted until the output looks acceptable. Sometimes a principle is invoked to constrain the range of weights assigned to the criteria. What principle, which weights, and who decides? Often, it's just up to the engineer.

One reason to be skeptical of this approach is that it seems to put the cart before the horse. We might naturally think that a particular decision output should be driven by the relations among the instantiations by each option of the covering criteria. In determining the best person to

hire, for example, we might determine how each candidate fares with respect to loyalty, productivity, and team spiritedness, and then compare those different 'packages' of instantiations of each covering criterion against one another. The general, abstract weightings of the covering criteria – of loyalty, productivity, and team spiritedness – are then determined by – but do not determine – the specific decision outputs across all possible cases. This is intuitively how we think our deliberative decisions should be made.

The second strategy, however, gets things the wrong way around. It infers the general, abstract weights of the covering criteria by testing which weights yield what look to be correct or intuitive outcomes across a small sample of cases and then generalizes across new cases. There is, however, no guarantee that a set of assigned covering weights that lead to attractive decision outputs in a small sample of cases will yield similarly attractive decision outputs across all cases. Indeed, axiologists aware of the context sensitivity of value would argue that there is good reason to think that this strategy will lead to radical value misalignment. We need an approach to AI that captures the context sensitivity of values.

There is a deeper problem. We now come to Second Mistake about human values made by current AI design. Current strategies make a mistake about the structure of values involved in tradeoffs. They assume:

> *Trichotomy:* A tradeoff (or comparison) between two items – e.g., values, value bearers, or aspects of a single value – must be in terms of one of three basic relations: the one item is better than the other, worse than it, or they are equally good. Otherwise the items are incomparable, and no (rational) tradeoff is possible.

Trichotomy maintains that the conceptual space of comparability between two items with respect to any value is filled by the trichotomy of relations 'better than,' 'worse than,' and 'equally good'. If none of the trichotomy of relations holds, the items are incomparable with respect to that value and a choice between them is no longer within the scope of rationality. You can nonrationally plump for one over another, but don't think your choice can be justified.[30]

According to *Trichotomy*, tradeoffs can be modelled by a balance scale. If you want your AI to achieve the right tradeoff between efficiency and lack of bias, you will need to, as it were, put efficiency on one side of the balance scale and lack of bias on the other. There are only three possible outcomes: one side of the scale goes down (efficiency wins by some degree), it goes up (lack of bias wins to some degree), or the two sides are evenly balanced (efficiency and lack of

bias are equally important). But *Trichotomy* is mistaken. To see why, we need to take a philosophical detour into the structure of value.

III.

Hard Cases and the Structure of Value

Suppose you must choose between putting your year-end bonus toward helping the needy and getting your child the iPad she so desperately wants. Or between pursuing a career in philosophy and one in computer science or giving to Oxfam and donating instead to the Against Malaria Foundation (AMF). Fill out the details of each option so that neither is better than the other. Does it follow that they are equally good? A small but definite improvement in one of them – say, provision of 100 extra mosquito nets – would not settle the matter, which it would have to if they were equally good. Does it follow that they cannot be compared? If they can't be compared, rational choice between them would be precluded. But such cases are common in the course of rational decision-making.

These are *hard cases.* How should we understand them?

Leading technologists often assume that hard cases are cases of uncertainty.[31] Human ignorance is one of the most salient features of human existence so it is natural to reach for uncertainty, say, about a human's evaluative aim or facts about the world, as an explanation of a machine's misalignment or its failure to produce a definitive output that one option is at least as good as another. It might seem that cases that appear hard become easy with more information.

But uncertainty is not the right explanation of hard cases. Is it possible that we are sometimes faced with options that are qualitatively very different but in the same overall neighborhood of value so that even an omniscient god, surveying the options, would say that the case was hard. Imagine a god who can see your two possible futures – one as a tax accountant and the other as a lumberjack. Could you fill out two such possible futures so that they are qualitatively very different and yet neither is better than the other and a small improvement in one does not thereby make it better? Would two such futures be *impossible?* I doubt it. Or consider cases in which you have first person authority over the relative subjective value of things and nevertheless find it hard to say which is better. You have first person authority – you count as omniscient – on the matter of which of two teas tastes better to you right now – the

Green Jasmine or the African Rooibos. It's *possible,* rationally speaking, for you to judge that neither tastes better to you right now, and furthermore to judge that a small but definitive improvement in tastiness in one doesn't thereby settle the matter. You have all the information you need on your taste buds and yet you can rationally judge that neither is better than the other and nor are they equally tasty. They just taste really different.[32] So, we might surmise (unless we have a trichotomous ax to grind) that there are some hard cases that are not due to uncertainty.

Of course the cases over which we have first person authority tend to be rather trivial, turning on our purely subjective judgments in the moment. But if there are hard choices here, why not elsewhere? We should not be misled into thinking that uncertainty, which surely accompanies many of our most important hard cases, is what *makes* the case hard. If this is right, then hard cases are not about us – they are not about what we don't know. They are about how our options relate with respect to what matters in the choice between them. Neither option is better than the other and nor are they equally good. And yet they are options between which it seems perfectly intuitive that there is a rational choice to be had.[33]

Other technologists think that hard cases are cases of indeterminacy or vagueness.[34] For example, sometimes an algorithm will have a tough time classifying an image as a bus because it is a vague whether what is depicted is a bus.

Similarly, an AI could find it difficult to determine whether one option is better than another because of vagueness in the linguistic data on which it operates.

But indeterminacy is not a fitting explanation for hard cases. The difficulty in classifying some things as a bus can be resolved by arbitrarily stipulating a precise version of the concept that settles the matter of whether it is a bus. The difficulty of determining whether a child should be put into foster care, however, is not to be settled by flipping a coin between different more

precise versions of the concept of 'a child's best interest'. In general, the questions that mark hard cases, such as whether to hire someone, grant bail, swerve and hit one to avoid hitting five, are not questions that can be appropriately resolved simply by arbitrarily tightening up our concepts. These are substantive matters not to be resolved through linguistic stipulation or by a coin flip, strategies that are always intrinsically permissible to resolve cases of indeterminacy.[35]

I have argued elsewhere that recognizing hard cases means rejecting *Trichotomy.* We should instead adopt *Tetrachotomy*, according to which there are four, not three, basic ways in things can be related by value and thus four ways tradeoffs can be made with respect to evaluative criteria: one thing is better than the other, worse than it, they are equally good, or they are *on a par*. In hard cases, items are *on a par*: they are comparable, but neither is better than the other and nor are they equally good.

What, then, is parity? For our purposes a gloss will do.[36] Two items are on a par with respect to some covering criteria when they are 'bi-directionally' related, qualitatively very different, and yet in the same 'neighborhood' overall with respect to the criteria. This needs some unpacking. Two items are bi-directionally related if one is better than the other with respect to some of the relevant covering criteria but the other is better with respect to other relevant covering criteria. Bi-directionality holds in any case where there is no dominance or pareto-superiority across all the relevant covering criteria. Two items are qualitatively very different if one instantiates one quality (or set of qualities) of the covering criteria significantly while the other instantiates another, rather different, quality (or set of qualities) significantly. Being in the same neighborhood of value can be understood intuitively. Two student papers are in the same neighbourhood of 'goodness as an essay in metaphysics' when they both deserve a 'B'.[37]

Paradigmatic hard cases fit the bill. When choosing between careers, places to live, and kinds of life to lead, the hard choices are ones that have the above features. In choosing between two candidates, Arun and Bing, for a job, the choice will be hard if Arun is, say, highly productive, a poor problem-solver and not much of a team player, and Bing is less productive but a creative problem solver and terrific team player. They are qualitatively different, bi-directionally related, and – we can assume – in the same neighborhood of 'goodness as a hire' overall. They are on a par as potential hires. They present a hard case.

AI design as it currently exists makes no room for hard cases; it assumes *Trichotomy* about the structure of value. There are only three good machine outputs when the machine is

built to deliver outputs for choice: Choose option A! Choose option B! Flip a coin between them! If value – and normativity more generally – has a tetrachotomous structure and items can be on a par, then technologists who build AI on the assumption of *Trichotomy* build machines that do not reflect the truth of human values, viz., that there are hard cases. Cases that are hard will be treated by trichotomous machines as easy. Misalignment is built into such designs.

The philosopher and AI scientist Bryce Goodman argues that hard cases, which he understands in terms of parity, put a 'hard limit' on the use of AI; we can't use technology in domains where we face hard cases because current algorithmic design makes no room for them.[38] We draw a different lesson from the existence of hard cases. Let's try to redesign AI to allow for hard cases and thereby bring machines in better alignment with human values. We humans face many hard choices. Machines, if they are to align with our values, should too.

IV.

What's a Machine to Do in a Hard Case?

How should we respond in a hard case? How should a machine respond? It is worth noting that if hard cases are understood as cases of uncertainty, a rational response can always in principle be to get more information. But as we've already suggested, in the hard cases of interest, sometimes more information does not help because we or God already have all the information we need. If, instead, hard cases are understood as cases of indeterminacy, a rational response can always in principle be to flip a coin to determine a precise concept that settles the matter. More precisely, in cases of indeterminacy, it is always intrinsically permissible – that is, permissible based on how the items relate – to resolve the case arbitrarily.[39] (There can always be an *extrinsic* reason to flip a coin if, for example, you're in a hurry.) But it is never intrinsically permissible to flip a coin to settle hard cases about child foster care placement, prison sentencing or final examination marks for students.[40] So hard cases don't fit how technologists commonly interpret them.[41]

When options for choice are on a par, a rational agent can do one of two things: *commit* or *drift.* By committing to one of the options, or more precisely, to some feature of an option, say, the winsomeness of Harry, you can endow an option with will-based value it didn't have before. This is a metaphysical claim about how a certain activity of the will, namely, putting one's very self behind some consideration can, under certain conditions, ground or be that in virtue of which something has value it didn't have before.[42] Humans themselves can be sources

of value.[43] Suppose you face a hard choice about with whom to spend the rest of your life: Tom, Dick or Harry. They are on a par. If you commit to Harry – if you put your self behind leading a life together with him – it may now be true that Harry is best for you. You've resolved your hard choice between Tom, Dick and Harry by committing to Harry. Your commitment also changes the reasons and values you have – your 'normative landscape' – going forward. By committing to Harry, the evaluative luster of late nights out carousing with strange men in bars diminishes. Crucially, committing is just *something you do qua* rational agent. While there may be reasons to commit to this rather than that, those reasons cannot guide your commitments. In this way, commitments are acts of pure rational agency: they are putting one's very self behind something: *I stand here!* Finally, and perhaps most importantly, what you do in hard cases determines your rational identity, that is, the composite of all the things you have most reason to do throughout your life.[44] Through your commitment in hard cases, you can quite literally *make it true* that you have most reason to do one thing rather than another. You make yourself into the sort of person who has most reason to spend her life with Harry rather than with Tom or Dick. This normative power to craft our rational identities is, I believe, central to human rational agency and irreplaceable by machines.

But you are not rationally required to commit in hard choices. You can instead rationally respond to a hard choice by *drifting* into one of the options, that is, intentionally choosing it for reasons but without standing behind it or adding value to it. You might drift into spending your life with Harry. This means that Harry is not better than Tom and Dick but on a par with them. If one night at a licentious party Tom or Dick gives you a come hither look, your normative landscape will be very different from what it would have been had you committed to Harry. This is how what you do in hard choices determines what you have most reason to do or feel going forward.

The above is a thumbnail sketch of a philosophical view about hard choices and the role of commitment in being a rational agent. If something like this view is correct, then we can extrapolate how a machine, if it is to align with human values, should respond in hard cases. Indeed, if it is not correct and hard choices are, contrary to arguments I have offered here and in other work, correctly understood as either an epistemic or determinate failure of trichotomy, any underlying mathematical model that fits the proposed framework could still be a *sine qua non* in achieving alignment. This is because a tetrachotomous mathematics that gives numerical

expression to the idea that there can be hard choices is likely to be philosophically interpretable in a variety of ways.[45] But since I think the right interpretation of hard choices is in terms of parity, I will continue investigating how machines could be built that align with this truth about the structure of value.

Suppose we want to build a tetrachotomous hiring algorithm that is programmed to deliver the 'best candidate for hire' at our widget factory. How would we do it? First we determine our covering criteria – the considerations that will cover 'best hire' and only 'best hire'. Let's suppose that we determine that evaluative criteria V, W, and X will do the job. Now we feed candidates into the algorithm, which then ranks candidates tetrachotomously for the position. Suppose it finds two top-ranked candidates, Arun and Bing, the choice between which the machine labels as 'hard'. What happens next? At this point the machine might send a message to the hiring committee indicating that Arun and Bing are the top candidates that are on a par. The hiring committee might then convene and review the dossiers of both candidates along with information about the candidates produced by the machine. They might discover that Arun is predicted to be highly productive and better than Bing with respect to criterion V, 'productivity', while Bing is predicted to be a highly creative team player and thus better than Arun on criterion W, 'creativity as a team member.' Perhaps they are on a par with respect to criterion X, 'loyalty.'

After debate and discussion, the committee might *commit* to productivity over creativity on behalf of the firm.[46] That is, they now 'stand behind' the importance of productivity as a hiring committee and see the normative landscape of that and competing values accordingly going forward. They help create the rational identity of the firm through this commitment. With the commitment in place, they might then send a response to the machine, asking it to make the minimal change required to the 'productivity' criterion, V, so that overall, Arun is now better than Bing, keeping fixed all other rankings already delivered.[47] This change in the importance of productivity then affects the machine's processing of other candidates going forward.[48]

Critically, human input of this sort cannot be replaced by machine input. The human commitment must be *actual*. This is because the value conferred by human commitment metaphysically depends on the actual commitment being made; there is a fact of the matter as to whether you commit to Harry's winsomeness, and thus whether Harry has value that he didn't have before that commitment. Even if the machine can predict with perfect accuracy what a

human would commit to, if there is no actual human commitment, then any changes the machine makes to criterial weights on the basis of that prediction will immediately entail misalignment with human values.

Suppose, for instance, that the hiring committee fails to make a commitment in the hard case involving Arun and Bing. But it successfully dupes the machine into believing that it has committed to Arun's productivity. The machine will make the minimal adjustment to productivity so that now Arun is better than Bing overall.

This will result in two sorts of alignment errors. Some cases that would have been hard without this adjustment are no longer hard, and the machine will not flag them. For example, prior to the adjustment giving greater importance to productivity, two subsequent applicants, Charlie and Delicia would have been a hard case. After the adjustment, Charlie, who is predicted to be marginally more productive than Delicia, would be ranked by the machine as better. But by the committee's actual values, Charlie and Delicia would be a hard case. So, the mismatch between the committee's actual commitment or lack thereof, on the one hand, and the machine's adjustment of criterial importance, on the other, would make some items better than others even though according to our human values, they would be hard cases. The reverse also holds. If there is a mismatch between an adjustment in criterial importance and the committee's actual commitment or lack thereof, cases that are by the committee's lights easy will be flagged by the machine as hard. If the committee were to review materials concerning Ebo and Frances, they would see that Ebo is clearly better. But the machine flags up the case as hard because Frances' marginal advantage in productivity now has greater weight than it did before the committee considered the case of Arun and Bing and lied about its commitment to productivity. In short, for a machine to align with human values, the cases it considers hard must be cases that humans consider hard. For there to be a match in values, humans must actually make a commitment in such cases, and the machine must register that commitment for there to be value alignment going forward.[49]

Now suppose the hiring committee reviewing the hard case of Arun and Bing *drifts* in favor of Arun over Bing. They have resolved the hard choice but not on the grounds that Arun is better than Bing. Rather they intentionally choose to go for Arun on the basis of reasons that count in favor of his being a good hire but not on the basis of reasons that make him the best hire. This resolution can then be communicated to the machine. In this case, the computer

processes Arun and Bing as a hard case going forward but also assigns a 'parenthetical' betterness ranking of Arun over Bing. In this way, the machine's values will continue to match the committee's while still enabling the machine to deliver a hiring output.

Notice that how a machine makes determinations is path-dependent; which cases it will flag as hard depends on which cases it previously encountered and what humans have sent back by way of adjustment in hard cases. This is how things should be. Human values evolve as we make commitments in hard cases we encounter throughout our lives.

V.
The Parity Framework of AI Design

We now have all the pieces we need to introduce the Parity Framework of AI design. This framework avoids the two fundamental mistakes about human values embedded in current AI design. It might, therefore, provide a promising path to human-machine value alignment.

The framework has three main features.

First, it takes a *values-based* approach to AI design and thus avoids the First Mistake about human values made by current AI design. Instead of processing nonevaluative data, such as where a job candidate went to school, as proxy for being the best hire, it processes evaluative data in the form of evaluative facts about options. There are two ways this evaluative data might be generated, directly and indirectly. The fact that Arun is a productive worker might be a datum gathered directly from judgements that Arun is a productive worker made by reliable or expert speakers. It might also be generated indirectly, by nonevaluative proxies, in particular, by evaluative judgements made by Arun and others. For example, perhaps Arun has a history of posting on social media things like 'I love exceeding my production targets!' and his peers post things like 'Arun is a dweeb because he's always working and never comes out to do karaoke'. Unlike the role of evaluative judgments in the direct route to evaluative facts, evaluative judgments operating indirectly are not *observations* of an evaluative fact but are proxies for an evaluative fact. Does this use of nonevaluative proxies violate *Values Proxy?* No, because the nonevaluative proxies for evaluative facts are not themselves data that the machine processes in determining whom to hire. Instead, these nonevaluative proxies work in the background to help generate the evaluative data on which the algorithm does its computations. Does the fact the

Parity Framework uses the fact that there is such-and-such an evaluative judgment on the internet, which is itself a nonevaluative fact, violate *Values Proxy?* No, because the nonevaluative fact that someone like an expert has made an evaluative judgment scraped from the internet is not the fact that is processed; it is the content of the judgement – which is evaluative – that is the input that the machine processes.

Second, the Parity Framework makes room for hard cases as a distinctive positive or good machine output, thereby recognizing the tetrachotomous structure of value. It thus avoids the Second Mistake about human values embedded in current AI design. Recognition of hard choices in turn creates a novel place for humans to be in the loop. Human input is critical whenever a machine encounters a hard case, and that feedback then alters the algorithm going forward. This process reflects the evolutionary nature of human values.

Finally, it supports an accompanying Parity Model – a specific model of decision-making – that is expressed by a tetrachotomous decision theory that offers a replacement for standard expected utility theory and its variants. The numerical model recognizes parity while rendering value computationally tractable. The mathematics comes courtesy of Kit Fine, the AI symbolic model is being built by engineers Luigi Bonasi and Bruno Lacerda, and the LLM machine learning version is being built by Sian Gooding of Deep Mind, all as part of a UK AISI Safety Grant co-led by the roboticist Nick Hawes and myself. Fine's basic mathematical idea is to represent numerically not the value of something but the 'approximate differences' in value between things. There are different possible versions of the Parity Model, but one that I think holds promise because it is the most intuitive is a combined machine-learning/symbolic system. In the case of a hiring algorithm, for instance, once certain evaluative covering criteria are initially determined, values-based machine learning can provide tetrachotomous rankings of candidates according to each covering value. Each covering value is then assigned weights in the form of an interval that can represent parity relations or, as Fine is exploring, perhaps those weights are expressed as a quotient or ratio of importance relative to one another. A sketch of such a possible hiring algorithm using the Parity Model and a values-based Parity Framework for AI design can be found in Chang 2023.[50] The difficulties in building and implementing this model are legion.

We can give a high level summary of the main differences current AI design and the proposed Parity Model design as follows.

1. Traditional AI design uses nonevaluative proxies to achieve evaluative ends. Parity design works directly with values (as reported by or otherwise derived from evaluative judgments) to achieve evaluative ends.

2. Traditional AI design permits only three good machine outputs: Choose x over y! Choose y over x! and Flip a coin! Parity design allows for four good machine outputs: Choose x over y! Choose y over x! Flip a coin! And Hard Choice!

3. Traditional AI design puts the human in the loop primarily as an epistemic tool for determining which of the three pre-existing possible outputs is correct and thus assumes that value is a static quality to be discovered. Parity design puts the human in the loop in a distinctive place, at the junctures of hard choices, where human feedback may change the outputs of the algorithm going forward as a reflection of the evolution of human values.

* * *

We end by considering some implications widespread implementation of the Parity Framework and its accompanying Parity Model could have on our future with AI. One obvious upshot is that any AI system that makes room for hard cases will not be fully autonomous since human input will always in principle be required. A utopian (dystopian?) future in which we can sit back, relax and let AI make all the difficult determinations for us is thus foreclosed. On the Parity Model, for example, AI won't have autonomous authority over normative decisions in domains such as hiring, prison sentencing, foster care systems, student evaluations, and so on. We humans will still have to roll up our sleeves and address hard cases in those domains. But arguably, this is as it should be; AI should not be autonomous in domains shot through with hard cases. And, as we've suggested, this embedded human autonomy in the deployment of AI would reflect the truth about human values: that they evolve with our commitments. Moreover, putting humans in the loop in this way would help ensure our own human autonomy.

Relatedly, the Parity Model may make AI too inefficient to be useful in some applications. Since human intervention is both cumbersome and time-consuming, the approach may restrict the domains in which such AI can effectively operate.[51] But again, this is arguably how things should be. The more riddled a domain is with hard cases, the less suitable it is for AI, even if built to flag such cases. So maybe this bug really is a feature.

Third, value-based approaches to AI design point the way to what we might call 'small ai'. If a machine cannot achieve evaluative purposes by pursuing nonevaluative proxies across a wide range of circumstances, alignment requires us to eschew the use of large models – even value-based ones if they come to exist – across a wide range of circumstances. Instead, we need to create – in the spirit of Aristotle's insights about the context-sensitivity of values -- *small ai*, that is small models that apply to specific sets of circumstances and specific sets of covering values. How can we do that? It will require expertise from both axiologists and domain experts. Suppose we want to create a hiring algorithm to deliver the best candidate for a job. Instead of creating a large model that tries to account for all covering values that could be relevant to being a good hire across all circumstances, we devise small models that *carve at the normative joints.* That is, we ascertain which covering values are relevant for which types of hiring situations. The covering values relevant for being the best hire in a university department will be different from the covering values relevant for being the best hire at a widget factory, which will be different still for hiring in a hospital or in a bank on Wall Street. We need to create separate value-based small parity algorithms at each of these normative joints.

This need not be as onerous as it may sound. For one thing, existing algorithms already can start us off with an initial list of covering considerations relevant for hiring in different domains, and we can fine-tune and supplement them with axiological expertise about which sets of covering values carve at the normative joints. For another thing, it will be for the most part perfectly intuitive how to determine the boundaries or joints of a small ai; creating an algorithm that tries to find the 'best hire' that will cover either a factory floor manager in a widget factory or a philosophy professor in a university will not carve at the normative joints. We can, I believe, determine the boundaries of small ai to achieve value alignment fairly readily and intuitively, especially with domain expertise at hand. Value Alignment, then, is alignment with human values at every normative joint. Existing 'fragmented' models created for specific tasks is in the spirit of what small ai would require. But instead of creating specific models for specific tasks, small ai would create specific models that addressed problems with the same covering values suitable for a range of contexts. Small ai would therefore be more efficient, allowing specific tasks to be gathered so long as they belonged to the same 'normative joint'.

Fourth, the Parity Model forces us to rethink what we treat as 'noise.' Daniel Kahneman, Olivier Sibony, and Cass Sunstein usefully define 'noise' as undesirable variation in judgment.[52]

Many of the cases they identify as noise, for example in prison sentencing and child foster care placement, however, are arguably cases in which at least some variation is due not to noise but to the hardness of the case. The analogy they employ is the bullseye of a target. Sometimes confounding factors prevent judgements from hitting the bullseye. They rightly tell us that we need to discard noisy judgments and focus on bullseye judgements. But the Kahneman, Sibony, and Sunstein don't countenance the possibility of hard cases. They assume, for example, that when a court makes a foster care assignment, it must always do what it can to hit the bullseye: the best, right assignment. Anything that doesn't hit the bullseye needs to be discarded as noise. But choosing in which foster home to place a child when the covering value is doing what is in the best interests of the child can be a paradigmatically hard choice. There may be multiple bullseyes or best assignments that are on a par. Adopting the Parity Framework and the Parity Model allows us to avoid conflating cases in which technology fails to hit the target – noise – with cases in which there are multiple legitimate targets to hit – hard cases.

Finally, the proposed Parity Framework might help meet general, sometimes amorphous, unease about the march of technology. This unease is sometimes expressed in terms of rights: we have a human right to explanation, we have a human right to human-made decisions. Because the Parity Model builds human intervention into the very design of AI, it permits partial explanation of AI outputs especially where such explanation is most needed, namely, in hard cases. Because AI decisions are the product of both machine and human inputs, those decisions are, by their very nature, partly human.

Underlying the Parity Framework is a view of machines and humans as forming a kind of hybrid intelligence,[53] one that allows machines to deliver a determination – about hiring, sentencing, foster care, academic evaluation and beyond – but only with critical interventions by humans at the junctures of hard cases. As it happens, people appear to trust decisions made by machine-human hybrids more than they do decisions made by machines or by humans alone.[54] The Parity Model offers a distinctive way to build trustworthy – *because aligned* – hybrid intelligence through AI design.

ACKNOWLEDGEMENTS: …

ENDNOTES

[1] One strand of the Control Problem cannot be solved by solving the Alignment Problem, namely, that of abuse of AI by bad actors. But if we build AI so that it does in fact align with human values, that is arguably a significant step in deterring bad actors, assuming that aligned AI – like an 'aligned person' -- is more difficult 'to turn' than misaligned AI.

[2] https://www.whitecase.com/insight-our-thinking/ai-watch-global-regulatory-tracker-united-kingdom#:~:text=International%20organizations%20including%20the%20OECD,not%20seem%20to%20have%20worked.

[3] Biden's Executive Order also included 150 specific regulations designed to ensure responsible AI from start to finish. An analysis of them can be found at https://hai.stanford.edu/news/numbers-tracking-ai-executive-order. (Disclaimer: the analysis of the regulations cited here – and which I found most helpful – was made by parties with some influence on their creation). To this author, they appear thoughtful and sensible, but they had the effect of putting out of commission for use by the federal government any non-cured software, which no doubt ruffled some businesses' feathers. The five principles of the AI Bill of Rights were (https://www.whitehouse.gov/ostp/ai-bill-of-rights/:) 1) You should be protected from unsafe or ineffective systems; 2) You should not face discrimination by algorithms and systems should be used and designed in an equitable way; 3) You should be protected from abusive data practices via built-in protections and you should have agency over how data about you is used; 4) You should know that an automated system is being used and understand how and why it contributes to outcomes that impact you; and 5) You should be able to opt out, where appropriate, and have access to a person who can quickly consider and remedy problems you encounter. It is perhaps worth hazarding that if algorithms were redesigned in the ways proposed in this paper and certain other practices, such as data practices, improved, alignment with the values underwriting these five principles could be much improved if not achieved.

[4] https://www.gov.uk/government/publications/ai-safety-summit-2023-chairs-statement-2-november/chairs-summary-of-the-ai-safety-summit-2023-bletchley-park#:~:text=The%20Bletchley%20Declaration%20agreed%20an,responsible%20AI%20that%20is%20safe and https://www.gov.uk/government/publications/ai-safety-institute-overview/introducing-the-ai-safety-institute.

[5] https://www.instituteforgovernment.org.uk/explainer/artificial-intelligence-regulation#:~:text=The%20AI%20white%20paper%20has,different%20fields%20of%20potential%20application.

[6] http://www.cac.gov.cn/2023-10/24/c_1699806932316206.htm; see also https://www.eastasiaforum.org/2023/09/27/the-future-of-ai-policy-in-china/.

[7] https://legalnodes.com/article/global-ai-regulations-tracker.

[8] From the Prime Minister's Office, 10 Downing Street and The Rt Hon Rishi Sunak MP, Speech on AI, at the Royal Society, 26 October 2023 (Transcript of speech as delivered), https://www.gov.uk/government/speeches/prime-ministers-speech-on-ai-26-october-2023.

[9] https://www.thestar.com/business/2023/05/24/ai-guru-yoshua-bengio-says-regulation-too-slow-warns-of-existential-threats.html. For what it's worth, my own view is that because of the problems mentioned above, at this point in time, regulation should primarily focus on solving the cluster of safety issues having to do with abuse of AI by bad actors.

[10] [Discuss problems of learning: generalization, memory, transferring information, knowledge vs. understanding, lack of common sense. Discuss problems of reasoning: causal vs. correlative reasoning, stochastic parroting of LLMs. Explain LLMs and mapping problem as an example].

[11] There is a third group that thinks scaling up present development interventions will lead us to Artificial General Intelligence, in which case all bets for alignment and control are off. I doubt that without some qualitative shift in AI design that we can reach AGI, but this assumes that AGI must 'pass through' regular human intelligence, of which current development interventions seem to fall well short.

[12] Norbert Wiener, 'Some Moral and Technical Consequences of Automation: As machines learn they may develop unforeseen strategies at rates that baffle their programmers,' *Science* 131 (3410): 1355-1358, (1960), doi:10.1126/science.131.3410.1355. ISSN 0036-8075. PMID 17841602. Archived from the original on October 15, 2022. See also Brian Christian, *The Alignment Problem: How Can Artificial Intelligence Learn Human Values?* (London: Atlantic Books, 2020), and Peter Asaro, 'Roberto Cordeschi on Cybernetics and Autonomous Weapons: Reflections and Responses', *Paradigmi: Revista di critica filosofica*, Anno XXXIII, no. 3, Settembre-Dicembre, 2015, pp.83-107.

[13] Asaro, Eckersley, Russell, Christian.

[14] Sorg, Singh, Lewis 2010.

[15] A diagnosis may be in order. Why have technologists almost uniformly assumed, a la Weiner, that alignment is a matter of whether machine output satisfies our preferences? I suspect that computer scientists and engineers, queasy about and unfamiliar with the idea of 'human values', look over to the social sciences, and in particular, economics, for help in thinking about value. Some welfare economists use preferences as a proxy for an agent's well-being. This makes sense when the aim is to aggregate, on the basis of measurable observables, the well-being of individuals to guide policymaking. But with AI in hand, we can do better. We can build 'values-based' AI whose alignment is tested not against preferences but human values themselves. More on this below.

[16] A word about 'principles' as possible metrics or guides for alignment. Like preferences, principles are, I believe, not the right phenomenon by which to measure value alignment. This is despite a burgeoning literature that gives principles pride of place in thinking about alignment. See e.g., Iason Gabriel, 'Artificial Intelligence, Values & Alignment', Minds and Machines (2020) 30:411–437 https://doi.org/10.1007/s11023-020-09539-2. Consider the principle 'One should keep one's promises. To understand the content of this principle and to apply it appropriately, we need to answer specific questions such as whether the principle applies if the cost of keeping your promise is death or whether the self-sacrifice involved merely amounts to an excuse not to keep it, under what normative conditions does an undertaking not rise to the level of a promise, how to trade off the importance of fulfilling the obligation against the values to be achieved by breaking it, and so on. These are all questions that appeal to *values.* As with preferences, for principles to do their work, values must be brought into play. None of this is to say that principles are not critical in AI research and development. We need principles in order to forge international agreements about AI and more generally for AI governance (see e.g. James Manyika, '5 problems…check', UN Digital Compact; Fei, Fei Li, 'Now More Than Ever, AI needs a Governance Framework', *Financial Times,* Feb 8, 2025, https://www.ft.com/content/3861a30a-50fc-41c9-9780-b16626a0d2e8.). But for the purposes of alignment, principles are too general to do more than preclude the most extreme misaligned machines.

[17] By 'value' (and its cognates) I mean to include everything in the normative domain, including deontic criteria like obligation and rights, and evaluative excellences like scientific creativity. I

include processes as well as outputs as targets for alignment since we might well object to a machine that denigrates minorities on the way to unbiased outputs.

[18] Explain King Midas. Stuart Russell, *Human Compatible: AI and the Problem of Control,* (New York: Viking, 2019).

[19] https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G.

[20] https://www.emra.org/emresident/article/hotspotting. Thanks to James Guszcza for directing me to the case. Such covering failures are common. Another case: the World Bank's algorithm for determining recipients of aid in Jordan has been criticized by Human Rights Watch for failing to capture those most in need. Although the algorithm has 57 indicators, these are still too crude as proxies for the purpose of identifying citizens in greatest need. See https://www.technologyreview.com/2023/06/13/1074551/an-algorithm-intended-to-reduce-poverty-in-jordan-disqualifies-people-in-need/?truid=098727b8d002441461d6a670c6533972&utm_source=the_download&utm_medium=email&utm_campaign=the_download.unpaid.engagement&utm_term=&utm_content=08-18-2023&mc_cid=69f91c622a&mc_eid=237b42f754.

[21] https://worldpopulationreview.com/country-rankings/health-care-costs-by-country.

[22] Zaid Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan, 'Dissecting Racial Bias in Algorithm Used to Manage the Health of Populations', *Science* 366 (6464): 447-453, (2019), doi: 10.1126/science.aax2342.

[23] W.B. Gallie, 'Essentially Contested Concepts'. *Proceedings of the Aristotelian Society* 56: 167-198, (1955).

[24] Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, Stuart Russell, 'Cooperative Inverse Reinforcement Learning', 30th Conference on Neural Information Processing Systems (NIPS, 2016), Barcelona, Spain; Stuart Russell, 'Provably Beneficial Artificial Intelligence', Future of Life Institute (2017), https://people.eecs.berkeley.edu/~russell/papers/russell-bbvabook17-pbai.pdf; Stuart Russell, *Human Compatible, op cit.* With the human in the loop in at least one device, multi-agent reinforcement learning could also in principle be improved with respect to alignment.

[25] Cooperative inverse reinforcement learning (CIRL) builds upon 'inverse reinforcement learning' (IRL), which, like its progeny, leaves the machine's reward function uncertain. But instead of learning the reward function through an interactive game with the human, it learns the human's reward function by observing human behavior. The problem, however, is that IRL assumes human behavior is always optimal, but since it's not, it doesn't learn the correct reward function. See Stuart Russell, 'Learning Agents for Uncertain Environments' (extended abstract), *Proceedings of the Eleventh Annual Conference on Computational Learning Theory, COLT' 98, ACM,* New York, NY, pp. 101 (1255 103), (1998); Andrew Ng and Stuart Russell, 'Algorithms for Inverse Reinforcement Learning*', Proceedings of the Seventeenth International Conference on Machine Learning* (663 670), (2000).

[26] A further problem is that Russell understands *Alignment* in terms of satisfying preferences. As we've suggested, preferences are not the appropriate metric for value alignment; if this is correct, then the conceptual framework of CIRL would need to be recast in terms of values, which I believe it could be (but have yet to convince Russell of the necessity!).

[27] Brian Christian, *The Alignment Problem, op cit.*

[28] Note that even if a set of nonevaluative properties P are always disjunctively coextensive with an evaluative property V, we mere mortals could never know what they were and *which* such Ps were coextensive in a particular set of circumstances. An AI, however, might in principle be able

to keep track of what would likely be a nearly infinite disjunctive coextensive natural property for each evaluative property, which would then need to be indexed to particular circumstances. Sometime in the future, it might be possible for an AI to have knowledge of the world and know which circumstances obtain and are relevant to the problem at hand and then pursue the relevant non-evaluative disjuncts in the circumstances as adequate proxy for the evaluative property at issue. This would have to be done for every evaluative property and for tradeoffs between the various evaluative properties involved. A tall order, indeed! For now, we can safely reject Values Proxy. Those technologists who insist on thinking that a version of Values Proxy is acceptable so long as it is relativized to particular circumstances – that is 'fragmented' – should correspondingly modify their AI design to reflect this fact. As suggested below, 'fragmented' AI that assumes a more 'fragmented' version of Values Proxy will be less efficient that values-based AI.

[29] CIRL presents a third way. Like the proxy approach, CIRL folds the covering and tradeoff problems into a single problem to be solved in one fell swoop, viz., by the machine's learning the human's reward function. If *Values Proxy* is false, however, it cannot learn the human's reward function by clocking non-normative facts about her, such as her responses to certain questions.

[30] More precisely, your choice between incomparables cannot be intrinsically justified on rational grounds. There can always be extrinsic – usually contingent – reasons to plump one way rather than another in the face of incomparables.

[31] See e.g., Stuart Russell, *Human Compatible op cit.*

[32] A more careful version of this argument is in Chang, 'The Possibility of Parity', *Ethics*, 112: 659-688, 2002. For some empirical evidence that people make such judgments, see Michael Messerli and Kevin Reuter, 'Hard Cases of Comparison', *Philosophical Studies* 174 (9): 2227-2250, 2017.

[33] This is not to deny that some intuitively 'hard choices' are due to uncertainty. But the hard choices of importance are not. See Chang, 'Hard Choices', *Journal of the American Philosophical Association,* 92: 586-620, 2016.

[34] Roel Dobbe, Thomas Krendl Gilbert, Yonatan Mintz, 'Hard Choices in Artificial Intelligence, *Artificial Intelligence* 300: 103555, https://doi.org/10.1016/j.artint.2021.103555, 2021.

[35] […] and Miriam Schonfield.

[36] For an introduction to parity, see Ruth Chang, 'Introduction.' In *Incommensurability, Incomparability and Practical Reason,* ed. Ruth Chang, Cambridge: Harvard University Press, pp. 1-34. 1997; Ruth Chang, 'The Possibility of Parity' *op cit.* For an intuitive case for parity assuming comparability, see Ruth Chang, 'Parity: An Intuitive Case', commissioned by *Ratio* 29: 395-411, 2016. For an informal model of parity in terms of evaluative differences, see Ruth Chang, *Making Comparisons Count, op cit.;* Ruth Chang 'Parity: An Intuitive Case', *op cit.* For an informal supervaluational model of parity, see Wlodek Rabinowicz, 'Value Relations,' *Theoria* 74 (1):18-49, 2008; Wlodek Rabinowicz, 'Value Relations Revisited,' *Economics and Philosophy* 28 (2):133-164, 2012. For model that understands parity in terms of being 'almost better than', see Eric Carlson, 'Parity Demystified,' *Theoria* 76: 119-128, 2010. For a numerical model of parity that could replace standard trichotomous expected utility theory with a tetrachotomous theory and that could in principle be used in computer programming, see Kit Fine, 'A Numerical Model of Parity and Imprecision' (Ms). It is perhaps worth noting that parity is not a relation in which the evaluative difference between items is small (as described in McElfresh et al. 2021). Two qualitatively different items that are on a par may have an extremely large evaluative difference between them (see Ruth Chang, 'Parity, Interval Value, and Choice,'

Ethics, 114: 331-350, 2005. See also Fine (Ms) for a mathematical expression of parity in which the evaluative difference may be large.

[37] See my discussion of neighbourhoods in…. Compare Chrisoula Andreou's discussion of categories (great, good, okay, below average, poor) which she understands trichotomously […[]

[38] Bryce Goodman, 'Hard Choices and Hard Limits for Artificial Intelligence,' *Proceedings of 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES'21)*, May 19–21, 2021, Virtual Event. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3461702.3462539

[39] Things are more complicated for degree-of-truth theorists about vagueness. But I leave those theories – and epistemic theories for which it is controversial whether arbitrary stipulation is intrinsically appropriate – aside.

[40] There may also be extrinsic reasons *not* to flip a coin. Judges, for instance, have a duty not to make arbitrary determinations in legal cases whatever the underlying merits of the case and are regularly sanctioned if they do so. See e.g. https://www.nbcnews.com/id/wbna21600306.

[41] Note that Dobbe et al., who deem 'hard cases' as cases of vagueness, might be thought to overlook this distinction between intrinsic and extrinsic warrant for responses in hard cases. If it is vague whether Arun is a better hire than Bing, then it is always intrinsically permissible to flip a coin between them. Dobbe et al. go on to argue that hard cases should be resolved by democratic means. But the reasons to resolve vagueness by democratic means are extrinsic reasons. And it is unclear whether a democratically generated resolution of vagueness has legitimacy if, as an *intrinsic* matter, the case could *just as well* have been resolved through the flip of a coin. Indeed, it seems more plausible to think that democratic deliberation is a proper way to resolve cases in which options are on a par, when it is never intrinsically permissible to flip a coin between the options. So, I suggest that Dobbe means by 'hard cases' cases of parity.

[42] For two distinct arguments for this claim, see Chang, 'Grounding Practical Normativity: Going Hybrid,' *Philosophical Studies,* 164 (1): 163-187, 2013 (for a metaphysical argument) and 'Commitment, Reasons, and the Will', *Oxford Studies in Meatethics,* ed. Russ Shafer-Landau, Vol 8, 74-113, 2013 (for a normative argument).

[43] This general insight belongs to Christine Korsgaard and, in a different way, the existentialists. […]The way commitments – understood as putting one's very agency behind something – can generate value here is a third rendition of this general insight. Unlike Korsgaard…Unlike the existentialists…see my …]

[44] A useful contrast might be Christine Korsgaard's notion of a *self-conception.* One's rational identity is the in-fact conglomeration of what one has most reason to do throughout one's life, not how one thinks of oneself. So, for instance, I might conceive of myself as an iconoclastic, bohemian free spirit, but it could nevertheless be true that the arc of the things I have most reason to do throughout my life involve achieving creature comforts and conformity with social norms. My self-conception can be wholly and directly a product of self-delusion while my rational identity cannot.

[45] In fact, a tetrachotomous numerical model of parity – a tetrachotomous decision theory that could and in my view should replace expected utility theory – that is being developed by Kit Fine is best interpreted as understanding hard choices as cases of parity. See Kit Fine 'Parity and…' Ms. But if you meet parity with what David Lewis called 'the incredulous stare', then the Parity Framework is still of interest so long as you think rankings over options are not always complete. (Cf Don Regan, Cian Dorr et al, and… a lot of economists). You would have an uphill battle, philosophically speaking, however, showing how committing or drifting are two intrinsically rational responses in hard choices, so you would need at the very least to modify the

framework and the underlying mathematics that so aptly expresses the idea that hard choices are cases of parity.

[46] They might instead commit to the specific kind of loyalty one candidate has over another or any criterion or any particular instantiation of a criterion. Another, more complex version of the Parity Model might permit decision-makers to commit to 'extraneous' criteria not built into the programming of the purpose, such as difficult background circumstances or being a member of an underrepresented minority. Arguably, the idea that a committee can decide between candidates that are otherwise on a par by committing to the value of diversity or overcoming discrimination is enshrined in Finland's Non-Discrimination Act, which holds that "an individual belonging to an underrepresented group can be selected from among applicants that have roughly the same qualifications" (as reported by Otto Sahlgren and Arto Laitinen, 'Computing Apples and Oranges? Implications of Incommensurability for (Fair) Machine Learning", Proceedings of the Ethicomp, Conference paper, 2022, p 7., SOURCE-WORK-ID: 67a967f3-2f0b-4bfd-9aa7-6195fcd8848c *Part of* ISBN: 978-951-29-8989-8. Parity is *not* the same as rough equality, which is typically understood as a phenomenon in which there is an underlying trichotomous truth about how the options relate. If there were a trichotomous truth about how two candidates compared, appealing to some extrinsic factor would be hard to justify. I believe the Finns have in mind parity without knowing it; arguably they are interested in cases in which the candidates are qualitatively very different but in the same overall neighborhood of value. That is why the government could commit the polity to adding value to being a member of an underrepresented minority.

[47] In the Parity Model, the 'minimal change' required is expressed numerically by adjusting the numerical representation of the evaluative 'ratio' difference between productivity and team-spiritedness.

[48] This is not to say that the firm must always use this altered algorithm in subsequent hiring years. After five years of off-the-charts productivity but a glum breakroom, the firm might adjust the algorithm to reflect the importance of team-spiritedness.

[49] Strictly speaking, a machine that could predict with 100% accuracy human commitments (or driftings) needs to pause its processing only so that, as it were, humans can 'catch up' and make the actual commitments (or driftings) that determine the values involved in cases going forward. It might be wondered why we should care whether the machine processes our commitments without us even if this leads to value misalignment. Maybe this kind of misalignment isn't problematic. The answer, I believe, lies in the importance of our being the boss of our values. As moral philosophers have argued, it is important for each of us to come to our own moral conclusions, even if this means that we sometimes get things wrong. […]Similarly, it is important for us to be the actual generators of value in the evolution of our human values.

This point underscores a distinction between the 'robust' normative power we have when we commit and create value, on the one hand, and the 'weak' normative powers we have when we consent, forgive, promise, etc., and trigger value that was always conditionally there, on the other. In the case of consent, for instance, a doctor can permissible perform a procedure to save a patient's life even if the patient isn't in a position to consent but consents after the procedure. This is because there is value in having consent operate this way in such-and-such circumstances. Such weak normative powers generate reasons in virtue of values. Robust normative powers generate reasons in virtue of a metaphysical fact about the will and rational agency, not in virtue of the goodness of their operating in this way. For an argument that this is so, see my 'Commitment, Reasons and the Will,' …Thanks to Gideon Yaffe for raising a

question about this distinction which I discuss at great length in 'Do We Have Normative Powers?'…

[50] Chang, 'Human in the Loop!', in Dave Edmonds, ed., *AI Morality,* Oxford University Press, 2023.

[51] To maximize AI efficiency, we might design technology to 'skip over' hard cases that don't have significant effects on a determination. The trick, of course, is to find a balance between respecting important features of human values and making machines as useful as they can be. Building technology to recognize hard cases is the first step in achieving this balance.

[52] Daniel Kahneman, Cass Sunstein, and Olivier Sibony, *Noise: A Flaw in Human Judgement* (New York: Little Brown/Hachette, 2021).

[53] Ece Kamar, 'Hybrid Workplaces of the Future', *XRDS: Crossroads, The ACM Magazine for Students*, V23 (2): 22-25, (2016), https://doi.org/10.1145/3013488; Dominik Delleman, Philipp Ebel, Matthias Sollner, and Jan Marco, 'Hybrid Intelligence', *Leimeister Business Information Systems Engineering* 61(5):637–643 (2019), https://doi.org/10.1007/s12599-019-00595-2; and James Guszcza, David Danks, Craig Fox, Krisitan J. Hammond, Daniel E. Ho, Alex Imas, James Landay, Margaret Levi, Jennifer Logg, Rosalind Picard, Manish Raghavan, Allison Stanger, Zacharty Ugolnik, and Anita Wiliams Woolley, 'Hybrid Intelligence: A Paradigm for More Responsible Practice', (October 12, 2022). SSRN: https://ssrn.com/abstract=4301478; http://dx.doi.org/10.2139/ssrn.4301478.

[54] Kern, Christoph, Frederic Gerdon, Ruben L. Bach, Florian Keusch, Frauke Kreuter, 'Humans versus Machines: Who is Perceived to Decide Fairer? Experimental Evidence on Attitudes toward Automated Decision-Making', *Patterns* 3: 100591, (2022).