Forthcoming, Dave Edmonds, ed., *AI Morality*, Oxford: OUP.

Human in the Loop!

Ruth Chang

In the not-too-distant future, technology might help us decide which school to attend, what career path to pursue, and whether to marry and have children. It might also help our governments decide how much to tax its citizens, which regulations to impose on businesses, how to manage energy consumption and distribute healthcare resources, whether to go to war, and even how to manage deep political disagreements. We should, I think, welcome this decision-making assistance.

One way machines could help us is as a mere aid or tool, much like a calculator helps us decide what tip to leave at a restaurant. When deciding between a career in law and graphic design, for instance, a career App of the future might tell us the percentage of people, similar to us in psychological and cognitive profile, who have succeeded as lawyers compared to those who have succeeded in graphic design. We might then use this information as part of our deliberation about the pros and cons of pursuing each career. At the end of the day, however, we would make the decision in the old-fashioned, human way, with technological output as nothing more than an informational aid.

Today technology is typically deployed in just this way. Machines tell us that certain features highly correlate with being a successful employee; that having a certain profile is predictive of having exorbitant lifetime medical costs; that living in a certain postal code makes it more likely than not that a person will default on a loan; that granting bail to a defendant with such-and-such features involves a certain chance that they will commit another crime while free, and so on. What an employer, healthcare policy-maker, loan officer, or judge does with this machine-generated information, however, is up to them. They can treat it as settling the matter at hand, ignore it all together, or something in between.

Although this use of technology—as a mere aid to human decision-making— appears relatively tame, it has some significant downsides. The most notorious is that technologies—for example, facial recognition, machine learning, natural language models—give us biased outputs, reproducing our own prejudices and bigotries since they train on data provided by flawed human attitudes and behaviour. If we treat such outputs more like prophecies than the off-colour musings of a retrograde relative, we are likely to end up exacerbating unfairness at a systemic level. The problem of bias and unfairness, perhaps in conjunction with a general (and surprisingly resilient) fear that current technology is a stepping stone to an artificial general intelligence that could in principle obliterate us, has led some technologists to call for a red line to be drawn in the silicon: we should create technology that, at best, operates as a mere aid to but never as a substitute for human decision-making. Allowing machines to determine all by themselves what we humans should do—what career we should pursue, whom we should hire, whether we should grant a defendant bail, etc.—crosses a line and should be strictly prohibited.

This attempt to ring-fence technology strikes me as unfortunate for two reasons. First, whatever its wisdom, it is doubtful that any government or organization could ensure—certainly not globally—that technology by and large operate as mere aids to, rather than substitutes for, human decision-making. Second, the horse has already bolted: technology that operates as a substitute for human decision-making

is already on the scene. Social media algorithms already decide what news we read and banking algorithms have for years determined whether we will qualify for a mortgage. We should, I believe, accept what appears to be inevitable: that technology that operates as a substitute for human decision-making—that is, technology that makes choices for us—will continue to be built and, indeed, proliferate. And so we should focus our energies on making such technology as useful and safe as it can be.

One natural route to building useful and safe decision-making machines involves aligning human and machine values. Alignment is arguably the most important open problem in AI.1 Today, the leading strategy for attempting to achieve alignment is to 'put the human in the loop' of machine processing. By requiring human input at critical junctures of machine processing, we can—so the hope goes—bring machine decision-making in line with human values.

This idea of putting the human in the loop has a variety of implementations. In autonomous weapons, for instance, humans are 'in the loop' in that they must initiate such weaponry; a machine is not allowed to decide by itself whether to bomb a village. Sometimes humans are in the loop when they have the power to abort a machine's output; if a human doesn't approve of a machine's decision to use a cluster bomb, it can abort that output (this is also called being 'on the loop'). There are more interesting ways humans can be in the loop, too. If a machine learning algorithm is uncertain about its reward function, it can observe or interact with humans and thereby—in principle—learn the human's reward function.2 In mixed machine learning rules-based programming, a human might input specific weights to be assigned to factors in decision-making. 3

This chapter sketches a novel way of putting the human in the loop of machine processing. The key to the proposal is to design technology so that the machine recognizes hard choices. Hard choices are between alternatives that can be compared, and yet neither is better than the other and nor are they equally good: they are on a par. Typically, alternatives are on a par when one is better in some respects, the other is better in other respects, and yet neither is at least as good as the other overall. They are qualitatively different from one another, and yet in the same neighbourhood of value overall. Should you have children or devote yourself to a career? Should you be a physician or architect? Live in the country or city? We humans face hard choices all the time. If machines are to both align with our values and make choices for us, they should face hard choices too.

Surprisingly, current technological design makes no room for hard choices. Indeed, the philosopher and AI expert Bryce Goodman has argued that the existence of hard choices in human life places a 'hard limit' on building decision-making AI.4 While some of the most sophisticated technological design makes room for uncertainty, incompleteness/incomparability, and indeterminacy, 5 hard choices are a distinct phenomenon.6 They require a new approach to AI design.

If you face uncertainty, incomparability, or indeterminacy, it is always intrinsically permissible to arbitrarily select between options. By 'intrinsically' I mean 'on the basis of how the options relate to one another', in contrast to 'extrinsic' bases, such as what would make you most popular or save you the most time. If your choice is shot through with uncertainty, it's as if you are choosing between two black boxes, and it's permissible to flip a coin between them. If your options cannot be compared, they are outside the scope of rational choice and, once again, it is permissible for you to randomly select between them. And if it's indeterminate which option you should choose because a relevant concept is vague, you are permitted arbitrarily to tighten up that concept one way rather than another to settle the matter.7 Thus, a machine confronting uncertainty, incomparability, or indeterminacy in its options would always be intrinsically permitted to flip a coin. But in hard choices it is never intrinsically permissible to flip a coin to settle which to choose.