Ethical Implementation of Artificial Intelligence to Select Embryos in *In Vitro Fertilization*

Primary disciplines: Artificial Intelligence, Ethics, Reproductive Medicine

Abstract

AI has the potential to revolutionize many areas of healthcare. Radiology, dermatology, and ophthalmology are some of the areas most likely to be impacted in the near future, and they have received significant attention from the broader research community. But AI techniques are now also starting to be used in *in vitro fertilization (IVF)*, in particular for selecting which embryos to transfer to the woman. The contribution of AI to IVF is potentially significant, but must be done carefully and transparently, as the ethical issues are significant, in part because this field involves creating new people.

We first give a brief introduction to IVF and review the use of AI for embryo selection. We discuss concerns with the interpretation of the reported results from scientific and practical perspectives. We then consider the broader ethical issues involved. We discuss in detail the problems that result from the use of black-box methods in this context and advocate strongly for the use of *interpretable* models. Importantly there have been no trials of clinical effectiveness, a problem in both the AI and IVF communities, and argue against premature implementation. Finally, we discuss ways for the broader AI community to become involved to ensure scientifically sound and ethically responsible development of AI in IVF.

Introduction

In vitro fertilization (IVF) is a clinical technique which has revolutionized the treatment of infertility. The process involves fertilizing the egg in a laboratory and replacing the resultant embryo into the uterus. Natural fertilization and conception is an inefficient process, with low chances of a live birth for any particular embryo. The solution both in nature and with medical treatment is to create many embryos, so that ultimately one will probably implant. In nature, the cost is time to pregnancy or, in the event of no embryos implanting, the pain of childlessness. In clinical practice, the cost is additionally measured in dollars and limited access to treatment. To increase the efficiency of clinical practice, much attention has been given to selecting the embryo that is most likely to implant. A recent innovation in the laboratory is time-lapse imaging of the embryo in culture over a number of days. This gives rise to thousands of visual data points and with it the promise of augmenting the embryo selection process with artificial intelligence (AI)-based models. In this paper, we provide an overview of the IVF process, review current approaches to using AI in embryo selection, discuss ethical issues of using AI in this specific field, and make proposals for the ethical implementation of this new technology. We finish with encouragement for AI researchers to collaborate with fertility clinicians to take this research forward in a meaningful and ethical way.

The *In Vitro Fertilization* (IVF) Process

Each year, millions of couples who suffer from infertility pin their hopes of starting or growing their family on IVF (European Society of Human Reproduction and Embryology 2020). Heavily criticized by many at first as an unethical human experiment (Fauser and Edwards 2005), the technique has become one of the most successful therapeutic innovations of the past half-century, leading to over 9 million babies born since the first IVF birth in 1978 (European Society of Human Reproduction and Embryology 2020). The limit of IVF's success, however, is reflected in the millions more whose hopes have not been fulfilled. Particularly for those with advancing age and comorbidities, but also for every couple who tries, success is not guaranteed. On average, across all age groups, the live birth rate per treatment cycle is 26.1% in the UK (Human Fertilisation and Embryology Authority 2020).

To maximize the chance of retrieving a good quality egg and subsequent embryo, women are given hormone treatment to stimulate multiple egg development, a process known as controlled ovarian hyperstimulation. The doctor

harvests a woman's eggs at egg collection, a procedure performed under sedation, in which they pass a needle under ultrasound guidance into an ovarian follicle to aspirate the follicular fluid which contains the egg. The embryologist inspects the follicular fluid under the microscope and identifies the egg with its surrounding cells. They then add sperm to fertilize the egg, and culture the resultant embryos in the laboratory for 2-6 days, depending on the clinical situation. Although the number of eggs collected can range from 0 to >40, a typical number of eggs collected would be around 10, of which typically 6-8 would fertilize, and about 2-4 would typically reach the blastocyst stage around day 5 or 6. The embryologist will then select 1 or 2 embryos for transfer to the uterus; the patient is given hormonal support; and approximately 2 weeks later, a pregnancy test confirms if the patient is pregnant or not. This is known as a "fresh" cycle. Any unused embryos thought to be viable are then frozen for use later in case the fresh cycle is unsuccessful, or if successful, for a future sibling (Centers for Disease Control and Prevention 2020).

Early embryo development at the preimplantation stage is a very dynamic process. Hours after fertilisation, two pronuclei are formed carrying DNA material contributed by the sperm and the egg. The pronuclear membrane breaks down shortly before the first cell division, leading to a 2-cell embryo. As cells continue to divide, they become more compact with increased cell-to-cell interaction from 3 days post-fertilization. On day 4, the embryo reaches the "morula" stage, where borders between cells become invisible. During the next 1 to 2 days, cells separate into 2 layers, with a growing cavity formed between them; at this point, the embryo is called a "blastocyst". The outer layer of cells, also known as the trophectoderm, will become the placenta following implantation, while the inner layer (inner cell mass) will become the fetus. Both layers hatch out of the shell around the embryo (zona pellucida) before implantation into the endometrium.

Traditional embryo selection is based on several snapshot observations of an embryo under a microscope, at specific time points during culture (Figures 1-6). Considering the dynamic nature of embryo development, the static nature of the information collected in this method limits the accuracy of embryo selection (Gardner et al. 2015). Recent clinical application of time-lapse videography has enabled additional novel measures for embryo selection (Liu et al. 2014, Liu et al. 2015). However, debate is still ongoing regarding the best approach of using such time-lapse images for embryo selection (Liu et al. 2020).

Embryo freezing has been revolutionized in the past 10 years, as laboratories have adopted vitrification (rapid freezing), with pregnancy and live birth rates now comparable between fresh and frozen transfers (Rienzi et al. 2017). As the age of the uterus is known to have little effect on live birth rates (Human Fertilisation and Embryology Authority 2018), if all the embryos are replaced, albeit one at a time, embryo selection would not affect pregnancy or live birth rates per *egg collection*, as ultimately all embryos will be given the chance to implant. However, not all couples persist with treatment even in the presence of remaining frozen embryos (Human Fertilisation and Embryology Authority 2020 and Centers for Disease Control and Prevention 2020). For some couples, therefore, maximizing their chance of a live birth at an earlier transfer could raise their overall chances of having a baby. Other benefits of improving the embryo selection process might include a shorter time and lower cost to achieve a pregnancy because of fewer embryo transfers (Sunkara et al. 2020), thereby enabling more couples to have their baby when they had planned, possibly affecting future family planning. Therefore, IVF clinics are keen to adopt new strategies for embryo selection that improve on current success rates.

Examples of developments to select embryos more likely to implant include (1) allowing embryos to self-deselect via extended culture in the laboratory (Gardner et al. 2000), or (2) adding extra genetic testing such as pre-implantation genetic testing for aneuploidy (PGT-A) which is controversial because of its invasive nature and diagnostic imperfection (Kemper et al. 2020). These tools increase the birth rate per transfer from about 25% to 60% (Gardner et al. 2015), despite safety and accuracy concerns over both approaches.

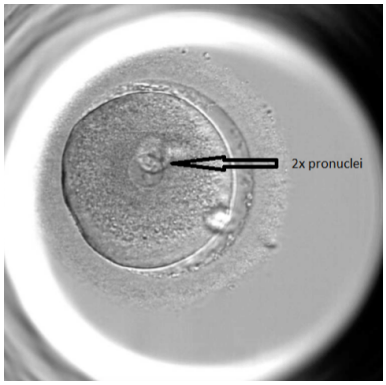

Figure 1: The embryo hours after fertilisation with 2 pronuclei

Figure 2: The 2-cell embryo

Figure 3: The 4-cell embryo

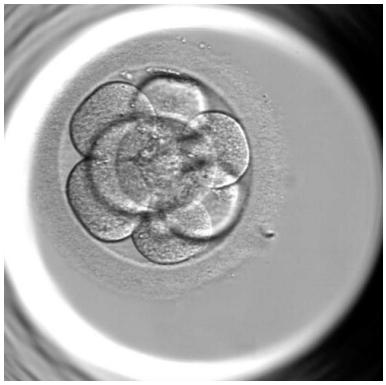

Figure 4: The 8-cell embryo

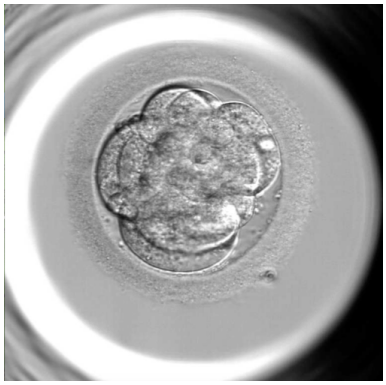

Figure 5: The “morula”

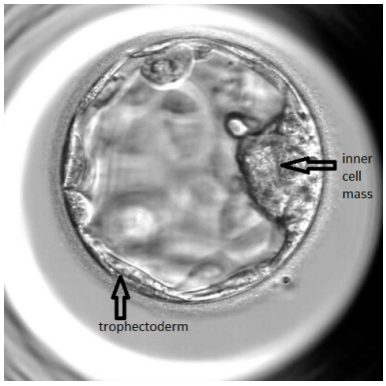

Figure 6: The “blastocyst” with trophectoderm and inner cell mass

AI as an Embryo Selection Tool to Improve the Success Rate Per Transfer

The application of AI in IVF has the potential to provide more objective, more rapid, and potentially more accurate evaluation of key steps in the IVF process, to make it more reproducible and repeatable when compared with a purely human approach (Rosenwaks 2020). In particular AI for embryo selection has attracted much interest, and potentially holds much promise (Fernandez et al. 2020).

The recent innovation of Time Lapse Imaging (TLI) in embryology offers the potential to generate vast quantities of data (Meseguer et al. 2011, Liu et al. 2016), which embryologists are still learning how to use (Liu et al. 2020). Computer Vision (CV) allows large amounts of image data to be automatically analysed by algorithms, and rapid recent advances in this field offer great promise to improve embryo selection.

More generally, the type of AI that can help embryo selection is Machine Learning (ML) – models that can

automatically learn and adapt as they are exposed to more data (whether images or other data). This is particularly useful when there is access to lots of data, but we do not know how to leverage it to make better predictions, or when we cannot manually process it all to generate meaningful knowledge. Potential variables include morphological features, such as cleavage of the embryo cells (blastomeres), fragmentation; morphokinetic features, which include time intervals between certain features; as well as considering confounding factors such as age of the woman or cause of infertility (ESHRE Working group on Time-lapse technology 2020).

Current State of Research on the Use of AI to Select Embryos

We searched MEDLINE and Embase for full-text studies evaluating AI to select embryos using the strategy included in the appendix. We checked the citations of papers we identified in the search for any publications we might have missed.

Studies evaluating AI for embryo selection make impressive accuracy claims for their ML models (Tran, Cooke and Illingworth 2019, Khosravi et al. 2019). One commonly reported performance measure is the receiver operating characteristic (ROC) curve which shows how a test's sensitivity and specificity correlate at different thresholds. The area under this curve (AUC) indicates the test's performance. An AUC >0.9 usually indicates outstanding performance, and the ML models from the studies cited above surpass this benchmark.

Studies which evaluate the efficacy of AI models for embryo selection do so for 2 types of outcomes; a) outcomes meaningful to the patient, such as a live birth or a fetal heartbeat positive pregnancy, or b) agreement with the existing standard, which in this case is assessment by embryologists.

Tran, Cooke and Illingworth's (2019) study belongs to the former category. They evaluated a model called IVY which rates how likely an embryo is to lead to a fetal heartbeat (FH) pregnancy on a confidence scale of 0 (definitely won't implant) to 1 (definitely will implant). Their ROC curve's AUC was 0.93. However, as Kan-Tor, Ben-Meir and Buxboim (2020) point out, the majority of the embryos on which the algorithm had been trained and tested were of such poor quality, that they would have been discarded in any event, thereby artificially inflating the AUC. As Kan-Tor, Ben-Meir and Buxboim explain, the clinical need is to identify the embryo with the highest chance of success among a set of embryos that appear to be potentially viable, and not from embryos which embryologists readily discard.

Khosravi et al.'s study on the other hand belongs to the latter category. They categorized their embryos into 3 groups – good-, fair- and poor-quality embryos according to a consensus of multiple embryologists. They then evaluated their AI algorithm's ability to identify the good- and the poor-quality embryos (but not the fair-quality embryos); for this task the algorithm achieved 96.94% accuracy (Khosravi et al.). This was better than the performance of individual embryologists. However, broad categorisations into "good" or "poor" quality are of limited benefit when trying to find the best embryo in a group of similar quality embryos.

The above analyses of Khosravi et al. and Tran et al.'s studies demonstrate the importance of understanding exactly how researchers test their algorithms before drawing conclusions from headline statistics.

Miyagi et al.'s team are the only ones who have used live birth as their outcome. They find an AUC much lower for their model (0.661) which was itself not significantly different from that of a prediction algorithm based on logistic regression analysis (0.713) which included features independent of the embryo (such as mother's age) as well as features dependent on the embryo (such as embryo diameter). Therefore, their ML model may perform on par with a potentially useful logistic regression algorithm in predicting a truly meaningful outcome – live birth – which might be of clinical benefit.

These studies are important steps to investigate efficacy (the ability to produce a specified outcome in experimental circumstances), to develop the tool, and establish proof of principle. However, they are only a prelude to testing in the clinic. When Curchoe et al. (2020) reviewed how the results of AI studies in reproductive medicine relate to real-life clinical practice, they highlighted four pitfalls that are common throughout the literature: small sample sizes, imbalanced datasets, non-generalisable settings and limited performance metrics.

Furthermore, to date, no AI studies for embryo selection have evaluated clinical effectiveness using the randomized controlled trial (RCT). The lack of RCTs appears to be typical of much of AI in medicine (Nagendran et al. 2020). The problem of lack of evidence before implementation is compounded by the IVF industry which is notorious for aggressively marketing unproven clinical and laboratory "add-ons" (Afnan, Khan and Mol 2020, Wilkinson et al. 2019). Furthermore, clinicians who do not have expertise in AI will find it difficult to critically navigate the literature which contains unfamiliar concepts and terminology.

Uninterpretable ("black box") machine learning models are either too complicated for any human to understand, or they are proprietary – in which case, comprehension of such a model is not possible for outsiders (Rudin 2019). Many studies in this field evaluate neural networks (Dirvanauskas et al. 2019) that are not interpretable, and not designed to be interpretable. Other approaches use interpretable features (whether they are labeled manually by doctors or labeled by neural networks whose output can be manually verified) but combine them in uninterpretable ways, such as using

principal components analysis (PCA) pre-processing (which forces a dependence on all variables) followed by a machine learning method such as a neural network or random forest (Chavez-Badiola et al. 2020, Milewski et al. 2017, Leahy et al. 2020). The work of Leahy et al. is interesting because the model is decomposable into separate computer vision models that each extract different information that can be checked by an embryologist. These separate models are combined into an uninterpretable neural network model to form the final prediction. Leahy et al.'s model is close to what we will recommend; if their final combined model was more interpretable, then each piece of the system would be either directly checkable by an embryologist for correctness or built as an interpretable model. A third category of studies use fully interpretable features, but use older techniques that are not particularly accurate and do not explicitly optimize for interpretability (e.g., the models are not sparse). These works generally do not apply any computer vision techniques, relying instead on humans to estimate measurements from the embryo images. Examples include the works of Raef, Maleki and Ferdousi (2019) and Morales et al. (2008), who created interpretable hand-calculated features and applied a variety of classical machine learning algorithms to them. The opaqueness or 'black box' nature of AI models is problematic for two main types of reasons: ethical, and epistemic.

Ethical Concerns with Black Box AI Models

Failure to Perform Randomized Controlled Trials

The most important ethical issue facing the adoption of AI assisted IVF is the need for careful randomized controlled trials against best current approaches. No matter how promising a new intervention appears to be, the gold standard for evaluation is the randomized controlled trial. Failing to do such trials risks harming patients, as does failing to perform systematic reviews of existing evidence and publish negative results (Savulescu, Chalmers and Blunt 1996).

Compromised Shared Decision Making

The second concern centers on the use of opaque AI models, potentially compromising shared decision-making and patient-centered medicine more broadly. Over the past few decades, the accepted model of clinical practice has shifted from a paternalistic one, where the clinician's opinion and recommendation is simply accepted by the patient, to one of shared decision-making where this power and informational asymmetry is reduced to the benefit of patient autonomy (Charles, Gafni and Whelan 1997). Clinical AI models which are opaque (and so medical explanations for a model's recommendation are inaccessible) compromise this shared decision-making due to the inability of the clinician and the patient to understand the model's decision (Bjerring, and Busch 2020). While there have been some counter-arguments raised as to whether shared decision-making is truly compromised by opaque AI models (Mishra, Savulescu and Giubilini [forthcoming]), application of AI in embryo selection should be guided by an awareness of such potential dangers. It will be important to fully explain what is known about how the AI model comes to a "decision" (nature and size of dataset, reasons for confidence in prediction, possible alternative lines of justification, etc.), and further examine how interactions between clinicians and patients may change, both at the point of embryo selection as well as at the point of implantation failure. Even if only one embryo is transferred, clinicians should explain the basis of this decision, whether it is clinical or AI-assisted. If information that is traditionally conveyed to the patient as to why a particular embryo is selected – for example the number and symmetry of the cells, or if the cells are fragmented, and therefore what the chances of implantation are, and why implantation may fail – are no longer accessible, shared decision-making might indeed be compromised. Existing measures of shared decision-making and decision quality, such as the Decision Conflict Scale (Garvelink et al. 2019), the OPTION Scale (Elwyn et al. 2003), and the SURE Test (Légaré et al. 2010) (among other patient-reported measures) can be used to guide such an evaluation.

It is important, however, not to overstate this concern. Firstly, AI-assisted decision-making should be compared to the status quo. Current expert judgment is based on biologically meaningful measures, which although are more broadly communicable than decisions of opaque models, are not very accurate for predicting a live birth. AI-assisted decision making may not be worse. More importantly, autonomy requires understanding information relevant and meaningful to one's values. Knowing the basis of a prediction (cleavage rate, symmetry, etc) is not relevant: what is relevant are the risks, side-effects and benefits, and the confidence attached to these assessments.

Misrepresentation of Patient-Values

Another ethical issue concerns potential harms from a misrepresentation of patient-values in the decision process. For example, there are reported differences between early morphokinetic profiles between male and female embryos (Wang et al. 2018, Tarín et al. 2014) (and other traits might be similarly differentially represented at this early stage). Models for embryo selection run the risk of systematically selecting for these traits if they are perceived by the model to be correlated with implantation success. For example, if a patient prefers that sex be randomly selected, this model may run counter to those values. If such models are opaque, this systematic favoring of particular traits may not

be detected at the time of decision-making. Further, if some of these traits are ethically salient ones for the patient, then this creates a scenario where the patient's values may not be sufficiently represented to guide the decision-making process for embryo selection. Such concerns have also been raised for other clinical models (McDougall 2019), calling for the design of such systems to be 'value-flexible' so that in clinical settings, both clinicians and patients are (1) aware of what metrics are driving a model's recommendations (either directly or as a proxy for some other medical fact/trait), and (2) able to appropriately reflect the patient's values in the decision process either directly through the model, or in subsequently adjusting the recommendation.

Again, it is important not to overstate this concern. The patient's own values could be inserted into AI algorithms (e.g., preference for sex and other non-disease characteristics) and AI might bring to the surface the importance of these values in decision making. Of course, valuing and selecting non-disease traits (such as sex or intelligence) raises the debate around designer babies, but some have argued that such selection is permissible (Agar 2004) or even a moral obligation when it relates to the well-being of a future child (Savulescu 2001; Savulescu and Kahane 2009; Savulescu and Kahane 2016).

Health and Well-Being of Future Children

Such potential biasing of AI-selection might also have impacts on the health or well-being of future children. For example, it is possible that some disadvantageous trait (such as increased risk of cancer or mental disorder) correlates with higher chance of implantation. However, this risk might be present unknowingly in ordinary clinical decision making. This also underscores the importance of clinical trials not merely measuring implantation or even healthy live birth but long term well-being of the child created by IVF through long term (decades) follow up.

Reproduction is also unique because selection determines who will come into existence. This creates the so-called "non-identity problem" which has spawned decades of unresolved philosophical debate, sparked by Derek Parfit (1984). Imagine Embryo A has a higher chance of implantation but unknowingly a higher chance of cancer later in life than embryo B. AI selects A. A is born but gets cancer at the age of 30. Was A harmed by the decision to select A rather than B? No, a different person (B) would have been selected. Provided that the disadvantageous trait or genes do not make A's life so bad as to have been not worth living, then A cannot be harmed by selection. On this ground, greater risks can be taken in embryo selection than with interventions on a specific embryo (such as A) which do risk harm to a specific individual (Savulescu et al. 2006). Nonetheless, some have argued that parents (and clinicians) still have a moral obligation to select the embryo with the best chance of the best life (Savulescu 2001, Savulescu and Kahane 2009; Savulescu and Kahane 2016).

Implications of Disvaluing Disability

There is a general problem with embryo selection raised by disability activists: any kind of selection based on predicted health or well-being discriminates against the disabled and expresses a negative message about the value of their lives - the expressivist objection (Buchanan et al. 2000). For example, screening for Down Syndrome has been said to express a negative view about the value of people with Down Syndrome (Hofmann 2017). This applies not only to AI selection but clinical selection and there are numerous responses (Buchanan et al. 2000). However, AI might considerably expand the scope of this objection: any trait which lowers chance of implantation might result in selection against that group, sex being an example we have discussed. The best response to these concerns would be to monitor such effects and ensure social responses that reinforce the equality of all people, including people with disabilities. Thus, rather than forgoing selection, it is better to ensure there are sufficient social resources so that all existing people have a reasonable chance of a good life (Savulescu and Kahane 2016).

Societal Implications of Bias in Embryo Selection

Successful AI models might be deployed at scale, and if such models systematically favor certain traits represented in early morphokinetic profiles, this might impact society. Even if would-be parents might not care about the sex of their future child and might be willing to accept a higher likelihood of one sex for a higher likelihood of implantation success, this will still have societal ramifications through a skewed population ratio. The scale of these ramifications will correlate with rates of IVF use in the future; the more individuals opt for IVF, the greater the impact. While such possibilities are at this stage mostly speculative, they represent a scale of impact that is significant and should therefore be considered.

Black Box Models Pose a Responsibility Gap

The final ethical issue concerns a potential erosion of ethical and legal accountability through the use of opaque AI models. If it is so determined that clinicians can't be held responsible for injuries sustained by the patient due to a reliance on opaque AI models, the responsibility for this class of errors would need to be borne by another agent. In the absence of institutionalized accountability mechanisms that hold other agents, such as model developers, responsible, this creates a 'responsibility gap' when it comes to the use of AI models.

The most straightforward case in which accountability is required would be repeated implantation failure or low success rates due to suboptimal embryo selection processes,

and/or injury being sustained by the patient as a result of implementation of a model recommendation (either to the mother through surgical complications or the child when he/she is born - wrongful life or birth). Under such circumstances, if the patient seeks an account of what happened or advances a charge of negligence against the clinician, the decision-making process needs to be explicable. Traditionally, if a charge of negligence is advanced, experts assess the clinician's decision-making process, and depending on whether they deem this to be medically reasonable the clinician is either acquitted or held culpable. If AI models used for embryo selection reason in uninterpretable ways, it is unclear how a court might evaluate the doctor's decision-making, and subsequently unclear how responsibility for injury may be adjudicated (Schönberger 2019, Price, Gerke and Cohen 2019).

Black box models may also have accountability implications for poor outcomes in research settings, for instance in randomized control trials. If an RCT of a black box fails and the model causes harm to the treatment group, it becomes similarly difficult to ascertain through existing accountability mechanisms who ought to be held responsible.

Epistemic Concerns with Black Box AI Models

There are technical challenges posed by black box or opaque systems; it is unclear how we might assess the reliability of the model's predictions, eliminate potentially confounding factors at the decision-making point, and assess to what extent the model's accuracy is representative in a given use-case. In a field where 'add-on' clinical offerings are already widespread despite inadequate evidence of effectiveness, this epistemic problem is especially troubling (Wilkinson et al 2019).

Black Box Models Create Information Asymmetries

The use of black box models creates an information asymmetry between the company selling the tool and the clinicians having to make daily decisions as to which embryo to transfer. Using such models would force the embryologist to abrogate decision making to programs they themselves do not understand. It is not possible to fully evaluate whether to trust these complex models without an understanding of their reasoning processes.

Confounders are Rampant

If we do not understand what a black box model is doing, it is entirely possible that its predictions are based on confounders that should not be used as predictors. Confounders are often difficult to detect and cause models not to generalize. When coupled with a poor choice for evaluation metric, the confounding might not be noticed.

Let us construct a simple example where an obvious confounder and a standard (but ill-chosen) evaluation metric provide a situation where a useless model would appear to be excellent. In this example, the confounder is the mother's age, and the metric is overall AUC (not the AUC for an individual couple). It is largely possible that the mother's age is a major factor in predictions; what if it were the sole factor, so that a model based on an image of the embryo is predictive of mother's age only? If the model were a mere proxy for age, it would be entirely useless in discriminating between embryos from the same couple, yet it may still score highly on AUC because age alone is predictive of success in implantation. Here there are two problems that combine to be worse than either alone: a mismatched evaluation metric, and an inscrutable model that does not reveal the problem with either the predictions or the metric.

Real-Time Error-Checking is Harder with Black Box Models

The two problems discussed above (information asymmetry and the possibility of confounders) lead to a third problem, namely difficulty of error-checking the model in real-time as it makes predictions in the clinic. We would want the clinician to be able to determine whether the model is reasoning in a way that is obviously wrong and catch new problems immediately should they arise. For instance, after a change in camera setting, an algorithm might suddenly start thinking that the shape of a current embryo looks like an embryo from the training set with a completely different shape. A clinician could catch that problem immediately if they knew the reasoning process of the model.

The Economics of "Buying In" to a Brittle Model Does Not Favor Clinicians or Patients

A potential consequence of the problems of information asymmetry and confounders listed above would be that black box model performance may be brittle to changes from the system it was trained on, and thus would likely be limited to the ecosystem in which it has been shown to work. This means that a clinic using this model would need to buy into that ecosystem, ovarian stimulation regimens, use of incubators and culture medium amongst other potential variables. This gives AI companies a great deal of *economic power* over clinics, potentially increasing the cost of treatment.

Overall Troubleshooting is Difficult for Black Box Models

If the model were more interpretable, it might be easier to troubleshoot broad problems in the model (beyond serious

issues that might be noticed in real-time usage). This includes ethical concerns such as racial or sex bias, as well as epistemic issues with accuracy or subtle confounding. If interpretability reveals flawed reasoning processes, the designer would be forced to alter the model to use correct reasoning, leading potentially to more robustness across ecosystems.

Interpretable ML as the Way Forward in Embryo Selection

An interpretable ML model is a predictive model that is constrained so that a human can better understand its reasoning process (Rudin 2019). Interpretable ML is a field that dates to the beginning of AI, back to the days of expert systems. The benefits of interpretable models are clear: by understanding the reasoning processes of predictive models, physicians can troubleshoot them and justify their decisions (to patients, other physicians, and during lawsuits). Physicians can combine the reasoning process of an interpretable model with information that is not in a database. They would not need to place blind trust in a black box model. And patient values (chance of disability, sex, single vs. double embryo transfer and chance of implantation) can be more easily accommodated by interpretable models.

Focusing on increasing the use of interpretable AI models is an elegant approach to both the epistemic and ethical concerns, by dispelling the opaqueness and allowing precise explanations of model predictions. For instance, models that are not opaque have an advantage because their use preserves existing mechanisms of accountability to a greater extent. For models that are opaque, revisions to such mechanisms are necessary. While this gap might eventually be filled by revisions in existing mechanisms of accountability, interpretable models can be argued to offer the option to preserve the status quo by allowing clinicians to understand model decisions better and thus retain the responsibility. This argument is far from resolved, but it is a promising reason to favor interpretable models over black box ones.

To the extent that there exists little or no trade-off between how interpretable a model is and how accurate it is when it comes to embryo selection, interpretable models are thus a promising solution. If it is the case that there is a salient difference in performance between interpretable and non-interpretable models, alternative solutions to both of the above epistemic and ethical concerns might have to be developed, so that we may benefit from the higher predictive accuracy of non-interpretable models. For now, there is no reason to believe that a salient performance difference between interpretable and non-interpretable models would exist. Interpretable models perform just as well even for benchmark datasets in computer vision. In fact, interpretable models are easier to troubleshoot (as domains change, as unusual cases arise, as racial bias cases need to be investigated), and thus lead to overall better performance of the model.

A major question in interpretable ML is what interpretability metric to use, as these metrics must (by definition of interpretability in that domain) be domain dependent. For computer vision for natural images, there have been major successful efforts by numerous groups of researchers to create interpretable neural networks that do not lose accuracy over their black box counterparts. These neural models go well beyond modeling only the "attention" of the network (that is, where the network is looking within an image), and are particularly useful for computer vision problems. Such interpretable neural networks could use different types of logical reasoning processes, including:

- Case-based reasoning (variations on k-nearest neighbors): In this case, the network would point out which parts of a test image are similar to prototypical past cases. The prototypical cases are chosen by the network along with the ways in which images are similar to each other (e.g., Chen et al 2019). One could envision an embryologist looking at a test image of an embryo, with an interpretable ML model pointing out how similar parts of it look to other prototypical known embryos whose outcome is known.
- Latent space disentanglement, where all information about a single concept (such as mother's age, or embryo size, density or color) is forced to travel through a single node of a network. Another way to say this is that each axis of a latent space (where an axis corresponds to activation of a node) represents a concept. This helps to understand information flow through the network (e.g., Chen, Bei and Rudin 2020). These types of disentangled models could potentially be useful for separating out the type of equipment, the age of the mother, and other pieces of information that might be embedded within the image of the embryo.
- Networks that are imbued with logical structure, such as probabilistic decision trees. By forcing the network to reason logically, we may be better able to understand its reasoning process (e.g., Wu and Song 2019, Li, Song and Wu 2020).

There are many challenges still in designing interpretable neural networks, particularly when the domain experts themselves do not know what constitutes interpretability; in other words, there are many directions for future research. Similar approaches to those discussed above apply to other data types, including sound signals or other types of medical images.

For problems involving categorical or real data ("tabular" data, rather than image data, time sequence data, or text data), interpretable machine learning models can also be developed. These models can potentially take the form of a medical scoring system, which means a small number of integer "point" values that sum, and translate into a risk (Ustun

and Rudin, 2019). For tabular data, neural networks and other forms of black box models do not seem to provide additional accuracy, which means that optimized medical scoring systems might be as accurate as one could get (depending on the dataset) (Rudin and Ustun 2018, Lou et al. 2013).

An interesting direction for future research is to combine interpretable neural networks for computer vision (to handle the visual data) with interpretable models for tabular data (rule lists or decision trees, for instance) to form a global interpretable model that handles these heterogeneous data types.

One key point that has emerged from past research is that as long as one can design the interpretability metric carefully to match the domain, interpretable models tend not to lose accuracy relative to their black box counterparts (Rudin 2019, Chen et al 2019). As far as we know, modern methods for interpretable ML have not yet been fully applied to the domain of IVF.

In Table 1, we summarize the advantages of interpretable AI models over black box models.

- Black box models might compromise shared decision making.
- Biases may go unchecked in black box models potentially leading to a misrepresentation of patient values, unintended health consequences for potential future people, and unintended consequences for society.
- Black-box models pose a responsibility gap**,** whereas accountability in interpretable models lies mainly with the human decision-maker.
- Black box models create information asymmetries*,* shifting power away from clinicians towards companies who create them. They prevent clinicians from effectively determining whether to trust their predictions. Interpretable models permit the decision of trust.
- Confounders are rampant and harder to detect in black box models.
- Real-time error-checking is much easier with interpretable models than black box models.
- Overall troubleshooting (e.g., for harmful bias) is difficult for black box models.
- Economics of "buying into" a brittle model in an expensive black box ecosystem does not favor clinicians or patients.
- Interpretable models do not seem to lose accuracy over their black box counterparts and might even perform better because they are easier to troubleshoot.

Table 1: Summary box of reasons why interpretability gains an advantage over black box models

Recommendations

Rigorous Evaluation with RCTs

Researchers must evaluate AI models to select embryos using the gold standard of RCTs against best clinical judgement or black box AI, if these have been deployed into practice or show promising results. Key outcomes for evaluation include time to live birth, number of embryo transfers before live birth and associated cost analysis, as well as live birth per egg collection, and health of the baby. Researchers should monitor the effects of the new technology with post-implementation surveillance.

Interpretable AI

Programmers should build interpretable machine learning models where biologically meaningful parameters guide embryo assessment, reducing the risk of hidden biases in algorithms causing unintended harms to society, permitting better troubleshooting, and better enabling clinicians to counsel their patients on the thinking underlying their treatment.

Regulatory Oversight for Interpretable AI

The importance of interpretability should be captured in mechanisms of regulatory oversight. Current regulatory approaches attempt to capture medical AI models as a type of medical device - they should further require either that AI model developers not produce black-box models if interpretable models are shown to have similar performance, or that any black-box model must come with the next-best interpretable model considered and trialed. Further, despite the fact that the field of assisted reproductive technology utilizes 'good practice' regulation for many advancements (such that violations are not legally punished), this would not suit the many risks of AI in embryo selection as outlined above. A 'hard' regulatory stance that promotes interpretable models would be a more advisable approach.

Access to Data and Code

Data and code used to create ML models should be made publicly accessible. This would enable reproducible research and the advancement of an exciting and important academic field. A high-quality public model would, at the very least, provide a performance baseline for other models.

Respect for Patient Privacy and Autonomy

Procedures should be put in place for securing patient privacy when data is shared, such as data anonymisation. All patients who use AI to select embryos should give fully informed consent, including knowledge of limitations and unknowns, use of data and images, and harms and benefits as shown by RCTs. They should be informed of how a model

arrives at a recommendation. Where possible, patient values should be inserted into the reasoning process of selection models.

Involving the Broader AI Community

Many young ML researchers are eager to get their hands on data to try out the latest techniques, and are passionate about using the technology to make the world a better place. Their participation should be encouraged. Currently, datasets for embryo selection are not broadly available. A naïve release of such data may do more harm than good, potentially inviting simplistic evaluations of ML techniques that fall prey to many of the criticisms we have discussed. Releasing a dataset and suggesting evaluation criteria for it which reflects actual practice, and takes ethical concerns into account, will require a broader discussion between embryologists, ethicists, and researchers in AI and statistics, and will also require addressing privacy concerns. This discussion ought to continue after the data are released. Nonetheless, allowing the broader AI community to see the data and get involved in their analysis will ensure that flawed and biased evaluations do not easily fly under the radar. It will also likely bring other important issues into the open that we have not yet recognized.

We summarize our recommendations in Table 2 below.

- Use of replicable, interpretable machine learning tools and data
- Well designed and conducted RCTs
- Post implementation surveillance
- Regulatory oversight for interpretable AI
- Funding for public institutions to transparently develop and evaluate machine learning models, and open access to code used in models
- Procedures for maintaining security of patient/embryo data whilst permitting ethical data sharing
- Fully informed consent to use AI
- Inclusion of patient values into AI programmes where possible
- Participation from the broader AI community

Table 2: Summary box of recommendations

Conclusion

Starting or growing a family is an immensely significant decision; technology which could help individuals who make that decision realize their goal would be invaluable. We see potential for AI in IVF to help more couples have children, earlier in their treatment, and at a lower cost. However, researchers, companies and clinics must ensure that the technology they promote or adopt brings real, measurable benefits to patients and, most importantly, does no harm. In this article, we highlighted limitations of current ML models and the studies which evaluate them, we drew specific attention to the ethical concerns that this technology could introduce in its current form, and suggested a path forward in terms of model design and evaluation. Most importantly, we hope to see interpretable machine learning models that clinicians could understand, troubleshoot and explain to their patients, rigorously evaluated with RCTs. We believe these are essential for creating tools which are fit for use for real individuals, hoping to start or grow a family, in the clinical setting.

Conflict of Interest

The authors declare no conflict of interest.

Appendices

Search strategy for full-text studies evaluating AI to select embryos in MEDLINE and Embase:

- MEDLINE: (exp Artificial Intelligence/ OR Artificial intelligence* OR AI OR exp Neural Networks, Computer/ OR Deep learning OR Neural network* OR machine learning OR support vector machine OR automatic classification) AND (exp Fertilization in Vitro/ OR IVF OR in vitro fertilization OR embryo*).
- Embase: (exp artificial intelligence/ OR artificial intelligence* OR AI OR exp machine learning/ or machine learning* OR exp artificial neural network/ OR neural network* OR deep learning* OR exp Deep Learning/ OR exp support vector machine/ OR support vector machine* OR automatic classification) AND (exp fertilization in vitro/ OR ivf OR embryo* OR in vitro fertilization OR exp in vitro fertilization/).

References

Afnan, M.A.M.; Khan, K.S.; And Mol, B.W. 2020. Generating Translatable Evidence to Improve Patient Care: The Contribution of Human Factors. *Reproductive Biomedicine Online* 41(3): 353–356. doi.org/10.1016/J.Rbmo.2020.04.025.

Agar, N. *Liberal Eugenics: In Defence of Human Enhancement*. Blackwell Publishing: Malden

Bjerring, J.C., And Busch, J. 2020. Artificial Intelligence and Patient-Centered Decision-Making. *Philosophy and Technology* doi.org/10.1007/S13347-019-00391-6.

Bronet, F.; Nogales, M.C.; Martinez, E.; Ariza, M.; Rubio, C.; Garcia-Velasco, J.A.; And Meseguer, M. 2015. Is There a Relationship Between Time-Lapse Parameters and Embryo Sex? *Fertility and Sterility* 103: 396–401 E392. doi.org/10.1016/J.Fertnstert.2014.10.050.

Buchanan, A.; Brock D.; Daniels, N.; And Wikler, D. 2000. *From Chance to Choice: From Genetics to Justice*. Cambridge: Cambridge University Press.

Centers for Disease Control and Prevention. 2020. 2018 Assisted Reproductive Technology Fertility Clinic Success Rates Report. Atlanta (GA): US Dept of Health and Human Services. Available at: https://www.cdc.gov/art/pdf/2018-report/ART-2018-Clinic-Report-Full.Pdf. Accessed January 2021.

Charles, C.; Gafni, A.; And Whelan, T. 1997. Shared Decision-Making in the Medical Encounter: What Does it Mean? (Or It Takes At Least Two to Tango). *Social Science & Medicine*. Mar;44(5):681–92. doi.org/10.1016/S0277-9536(96)00221-3.

Chavez-Badiola, A.; Flores-Saiffe Farias, A.; Mendizabal-Ruiz, G.; Garcia-Sanchez, R.; Drakeley, A.; and Garcia-Sandoval, J. 2020. Predicting Pregnancy Test Results After Embryo Transfer by Image Feature Extraction and Analysis Using Machine Learning. *Scientific Reports* 10(1). doi.org/10.1038/s41598-020-61357-9.

Chen, C.; Li, O.; Tao, C.; Barnett, A.; Su, J.; and Rudin, C. 2019. This Looks Like That: Deep Learning for Interpretable Image Recognition. In Proceedings of the 33rd Conference on Proceedings of Neural Information Processing Systems (NeurIPS).

Chen, Z.; Bei, Y.; and Rudin, C. 2020. Concept Whitening for Interpretable Image Recognition. *Nature Machine Intelligence* 12(2): 772–782. doi.org/10.1038/s42256-020-00265-z.

Curchoe, C.; Flores-Saiffe Farias, A.; Mendizabal-Ruiz, G.; and Chavez-Badiola, A. 2020. Evaluating Predictive Models in Reproductive Medicine. *Fertility and Sterility* 114(5): 921–926. doi.org/10.1016/j.fertnstert.2020.09.159.

Dirvanauskas, D.; Maskeliunas, R.; Raudonis, V.; and Damasevicius, R. 2019. Embryo Development Stage Prediction Algorithm for Automated Time Lapse Incubators. *Computer Methods and Programs in Biomedicine* 177: 161–174. doi.org/10.1016/j.cmpb.2019.05.027.

Elwyn, G.; Edwards, A.; Wensing, M.; Hood, K.; Atwell, C.; and Grol, R. 2003. Shared Decision Making: Developing the OPTION Scale for Measuring Patient Involvement. *BMJ Quality & Safety* 12:93–99. doi.org/10.1136/qhc.12.2.93.

European Society of Human Reproduction and Embryology. 2020. ART Factsheet. Accessed 26 January 2021. https://www.eshre.eu/Press-Room/Resources

ESHRE Working group on Time-lapse technology; Apter, S.; Ebner, T.; Freour, T.; Guns, Y.; Kovacic, B.; Le Clef, N.; Marques, M.; Meseguer, M.; Montjean, D.; Sfontouris, I.; Sturmey, R.; and Coticchio, G. 2020. Good Practice Recommendations for the Use of Time-Lapse Technology. *Human Reproduction Open* 2020(2): 1–26. doi.org/10.1093/hropen/hoaa008.

Fauser, B.C., and Edwards, R.G. 2005. The Early Days of IVF. *Human Reproduction Update* 11(5): 437–438. doi.org/10.1093/humupd/dmi026.

Fernandez, E.; Ferreira, A.; Cecílio, M.; Chéles, D.; de Souza, R.; Nogueira, M.; and Rocha, J. 2020. Artificial Intelligence in the IVF Laboratory: Overview Through the Application of Different Types of Algorithms for the Classification of Reproductive Data. *Journal of Assisted Reproduction and Genetics* 37(10): 2359–2376. doi.org/10.1007/s10815-020-01881-9.

Gardner, D.K.; Lane, M.; Stevens, J.; Schlenker, T.; and Schoolcraft, W.B. 2000. Blastocyst Score Affects Implantation and Pregnancy Outcome: Towards a Single Blastocyst Transfer. *Fertility and Sterility* 73: 1155–1158. doi.org/10.1016/s0015-0282(00)00518-5.

Gardner, D.K.; Meseguer, M.; Rubio, C.; and Treff N.R. 2015. Diagnosis of Human Preimplantation Embryo Viability. *Human Reproduction Update* 21: 727–747. doi.org/10.1093/humupd/dmu064.

Garvelink, M.M.; Boland, L.; Klein, K.; Nguyen, D.V.; Menear, M.; Bekker, H.L.; Eden, K.B.; LeBlanc, A.; O'Connor, A.M.; Stacey, D.; and Légaré, F. 2019. Decisional Conflict Scale Use over 20 Years: The Anniversary Review. *Medical Decision Making* 39(4):301–314. doi.org/10.1177/0272989X19851345.

Hofmann B. 2017. 'You Are Inferior!' Revisiting the Expressivist Argument. *Bioethics* 31: 505–514. doi.org/10.1111/bioe.1236584.

Huang, B.; Ren, X.; Zhu, L.; Wu, L.; Tan, H.; Guo, N.; Wei, Y.; Hu, J.; Liu, Q.; Chen, W.; Liu, J.; Li, D.; Liao, S.; and Jin, L. 2019. Is Differences in Embryo Morphokinetic Development Significantly Associated With Human Embryo Sex? *Biology of Reproduction* 100: 618–623. doi.org/10.1093/biolre/ioy229.

Human Fertilisation and Embryology Authority. 2020. Fertility Treatment 2018: Trends and Figures: Quality and Methodology Report. Available at: https://www.hfea.gov.uk/about-us/publications/research-and-data/fertility-treatment-2018-trends-and-figures/fertilitytreatment-2018-quality-and-methodology-report/. Accessed January 2021.

Kan-Tor, Y.; Ben-Meir, A.; and Buxboim, A. 2020. Can Deep Learning Automatically Predict Fetal Heart Pregnancy with Almost Perfect Accuracy? *Human Reproduction* 35(6): 1473. doi.org/10.1093/humrep/deaa083.

Kemper, J.M.; Wang, R.; Rolnik, D.L.; and Mol, B.W. 2020. Preimplantation Genetic Testing for Aneuploidy: Are We Examining the Correct Outcomes? *Human Reproduction* 35: 2408–2412. doi.org/10.1093/humrep/deaa224.

Khosravi, P.; Kazemi, E.; Zhan, Q.; Malmsten, J.; Toschi, M.; Zisimopoulos, P.; Sigaras, A.; Lavery, S.; Cooper, L.; Hickman, C.; Meseguer, M.; Rosenwaks, Z.; Elemento, O.; Zaninovic, N.; and Hajirasouliha, I. 2019. Deep Learning Enables Robust Assessment and Selection of Human Blastocysts After In Vitro Fertilization. *npj Digital Medicine* 2(1). doi.org/10.1038/s41746-019-0096-y.

Leahy, B.; Jang, W.; Yang, H.; Struyven, R.; Wei, D.; Sun, Z.; Lee, K.; Royston, C.; Cam, L.; Kalma, Y.; Azem, F.; Ben-Yosef, D.; Pfister, H.; and Needleman, D. 2020. Automated Measurements of Key Morphological Features of Human Embryos for IVF. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020* 12265: 25–35. doi.org/10.1007/978-3-030-59722-1_3.

Légaré, F.; Kearing, S.; Clay, K.; Gagnon, S.; D'Amours, D.; Rousseau, M.; and O'Connor, A. 2010. Are You SURE? Assessing Patient Decisional Conflict with a 4-Item Screening Test. *Canadian Family Physician* 56(8):e308–14.

Li, X.; Song, X.; and Wu, T. 2019. AOGNets: Compositional Grammatical Architectures for Deep Learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6213–6223. doi.org/10.1109/cvpr.2019.00638.

Lin, J.; Zhong, C.; Hu, D.; Rudin, C.; and Seltzer, M. 2020. Generalized and Scalable Optimal Sparse Decision Trees. In Proceedings of the International Conference on Machine Learning (ICML).

Liu, Y.; Chapple, V.; Feenan, K.; Roberts, P.; and Matson, P. 2015. Clinical Significance of Intercellular Contact at the Four-Cell Stage of Human Embryos, and the Use of Abnormal Cleavage Patterns to Identify Embryos with Low Implantation Potential: A Time-Lapse Study. *Fertility and Sterility* 103(6):1485–1491.e1. doi.org/10.1016/j.fertnstert.2015.03.017.

Liu, Y.; Chapple, V.; Feenan, K.; Roberts, P.; and Matson, P. 2016. Time-Lapse Deselection Model for Human Day 3 In Vitro

Fertilization Embryos: The Combination of Qualitative and Quantitative Measures of Embryo Growth. *Fertility and Sterility* 105: 656–662.e1. doi.org/10.1016/j.fertnstert.2015.11.003.

Liu, Y.; Chapple, V.; Roberts, P.; and Matson, P. 2014. Prevalence, Consequence, and Significance of Reverse Cleavage by Human Embryos Viewed with the Use of the Embryoscope Time-Lapse Video System. *Fertility and Sterility* 102(5): 1295–1300.e2. doi.org/10.1016/j.fertnstert.2014.07.1235.

Liu, Y.; Feenan, K.; Chapple, V.; and Matson, P. 2019. Assessing Efficacy of Day 3 Embryo Time-Lapse Algorithms Retrospectively: Impacts of Dataset Type and Confounding Factors. *Human fertility: journal of the British Fertility Society* 22: 182–190. doi.org/10.1080/14647273.2018.1425919.

Liu, Y.; Qi, F.; Matson, P.; Morbeck, D.E.; Mol, B.W.; Zhao, S.; and Afnan, M. 2020. Between-Laboratory Reproducibility of Time-Lapse Embryo Selection Using Qualitative and Quantitative Parameters: A Systematic Review and Meta-Analysis. *Journal of Assisted Reproduction and Genetics* 37: 1295–1302. doi.org/10.1007/s10815-020-01789-4.

Lou, Y.; Caruana, R.; Gehrke, J.; and Hooker, G. 2013. Accurate Intelligible Models with Pairwise Interactions. In Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pp 623–631. doi.org/10.1145/2487575.2487579

McDougall, R.J. 2019. Computer Knows Best? The Need for Value-Flexibility in Medical AI. *Journal of Medical Ethics* 45:156–160. doi.org/10.1136/medethics-2018-105118.

Meseguer, M.; Herrero, J.; Tejera, A.; Hilligsoe, K.M.; Ramsing, N.B.; and Remohi, J. 2011. The Use of Morphokinetics as a Predictor of Embryo Implantation. *Human Reproduction* 26: 2658–2671. doi.org/10.1093/humrep/der256.

Milewski, R.; Kuczyńska, A.; Stankiewicz, B.; and Kuczyński, W. 2017. How Much Information About Embryo Implantation Potential is Included in Morphokinetic Data? A Prediction Model Based on Artificial Neural Networks and Principal Component Analysis. *Advances in Medical Sciences*, 62(1): 202–206. doi.org/10.1016/j.advms.2017.02.001.

Mishra, A.; Savulescu, J.; and Giubilini A. Forthcoming. Ethics of Medical AI. *Oxford Handbook of Digital Ethics*

Morales, D.; Bengoetxea, E.; Larrañaga, P.; García, M.; Franco, Y.; Fresnada, M.; and Merino, M. 2008. Bayesian Classification for the Selection of In Vitro Human Embryos Using Morphological and Clinical Data. *Computer Methods and Programs in Biomedicine* 90(2):104–116. doi.org/10.1016/j.cmpb.2007.11.018.

Nagendran, M.; Chen, Y.; Lovejoy, C.; Gordon, A.; Komorowski, M.; Harvey, H.; Topol, E.; Ioannidis, J.; Collins, G.; and Maruthappu, M. 2020. Artificial Intelligence Versus Clinicians: Systematic Review of Design, Reporting Standards, and Claims of Deep Learning Studies. *British Medical Journal* m689. doi.org/10.1136/bmj.m689.

Parfit, D. 1984. *Reasons and Persons*. Oxford: Clarendon Press

Price, W.N.; Gerke, S.; and Cohen, I.G. 2019. Potential Liability for Physicians Using Artificial Intelligence. *Journal of the American Medical Association*. 322(18):1765–1766. doi.org/10.1001/jama.2019.15064.

Raef, B.; Maleki, M.; and Ferdousi, R. 2019. Computational Prediction of Implantation Outcome After Embryo Transfer. *Health Informatics Journal* 26(3): 1810–1826. doi.org/10.1177/1460458219892138.

Rienzi, L.; Gracia, C.; Maggiulli, R.; LaBarbera, A.R.; Kaser, D.J.; Ubaldi, F.M.; Vanderpoel, S.; and Racowsky, C. 2017. Oocyte, Embryo and Blastocyst Cryopreservation in ART: Systematic Review and Meta-Analysis Comparing Slow-Freezing Versus Vitrification to Produce Evidence for the Development of Global Guidance. *Human Reproduction Update* 23: 139–155. doi.org/10.1093/humupd/dmw038.

Rosenwaks, Z. 2020. Artificial Intelligence in Reproductive Medicine: A Fleeting Concept or the Wave of The Future? *Fertility and Sterility* 114(5): 905–907. doi.org/10.1016/j.fertnstert.2020.10.002.

Rudin, C. 2019. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence* 1: 206–215. doi.org/10.1038/s42256-019-0048-x.

Rudin, C., and Ustun, B. 2018. Optimized Scoring Systems: Toward Trust in Machine Learning For Healthcare and Criminal Justice. *INFORMS Journal on Applied Analytics* 48(5):449–466. doi.org/10.1287/inte.2018.0957.

Savulescu, J. 2001. Procreative Beneficence: Why We Should Select the Best Children. *Bioethics* 15 (5): 413–426. doi.org/10.1111/1467-8519.00251.

Savulescu, J.; Chalmers, I.; and Blunt, J. 1996. Are Research Ethics Committees Behaving Unethically? Some Suggestions for Improving Performance and Accountability. *British Medical Journal* 313: 1390–3. doi.org/10.1136/bmj.313.7069.1390.

Savulescu, J.; Hemsley, M.; Newson, A.; and Foddy, B. 2006. Behavioural Genetics: Why Eugenic Selection is Preferable to Enhancement. *Journal of Applied Philosophy*. 23(2): 157–171. doi.org/10.1111/j.1468-5930.2006.00336.x.

Savulescu, J., and Kahane, G. 2009. The Moral Obligation to Create Children with the Best Chance of the Best Life. *Bioethics* 23(5):274–290. doi.org/10.1111/j.1467-8519.2008.00687.x.

Savulescu, J., and Kahane, G. 2016. Understanding Procreative Beneficence: The Nature and Extent of the Moral Obligation to Have the Best Child. In *The Oxford Handbook of Reproductive Ethics* ed. L. Francis. Oxford: Oxford University Press. doi.org/10.1093/oxfordhb/9780199981878.013.26.

Schönberger D. 2019. Artificial Intelligence In Healthcare: A Critical Analysis of the Legal and Ethical Implications. *International Journal of Law and Information Technology* 27(2): 171–203. doi.org/10.1093/ijlit/eaz004.

Sunkara, S.K.; Zheng, W.; D'Hooghe, T.; Longobardi, S.; and Boivin, J. 2020 Time as an Outcome Measure in Fertility-Related Clinical Studies: Long-Awaited. *Human Reproduction* 35: 1732–1739. doi.org/10.1093/humrep/deaa138.

Tarín, J.J.; García-Pérez M.A.; Hermenegildo, C.; and Cano, A. 2014. Changes in Sex Ratio from Fertilization to Birth in Assisted-Reproductive-Treatment Cycles. *Reproductive Biology and Endocrinology* 12: 56. doi.org/10.1186/1477-7827-12-56.

Tran, D.; Cooke, S.; and Illingworth P.J. 2019. Deep Learning as a Predictive Tool for Fetal Heart Pregnancy Following Time-Lapse Incubation and Blastocyst Transfer. *Human Reproduction* 34(6): 1011–1018. doi.org/10.1093/humrep/dez064.

Ustun, B., and Rudin, C. 2019. Learning Optimized Risk Scores. *Journal of Machine Learning Research* 20(150):1–75

Wang, A.; Kort, J.; Behr, B.; and Westphal L.M. 2018. Euploidy in Relation to Blastocyst Sex and Morphology. *Journal of Assisted Reproduction and Genetics* 35: 1565–1572. doi.org/10.1007/s10815-018-1262-x.

Wilkinson, J.; Malpas, P.; Hammarberg, K.; Tsigdinos, P.M.; Lensen, S., Jackson, E.; Harper, J.; and Mol, B.W. 2019. Do A La Carte Menus Serve Infertility Patients? The Ethics and Regulation of In Vitro Fertility Add-Ons. *Fertility and Sterility* 112(6): 973–977. doi.org/10.1016/j.fertnstert.2019.09.028.

Wu, T., and Song, X. 2019. Towards Interpretable Object Detection by Unfolding Latent Structures. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR). doi.org/10.1109/iccv.2019.00613.